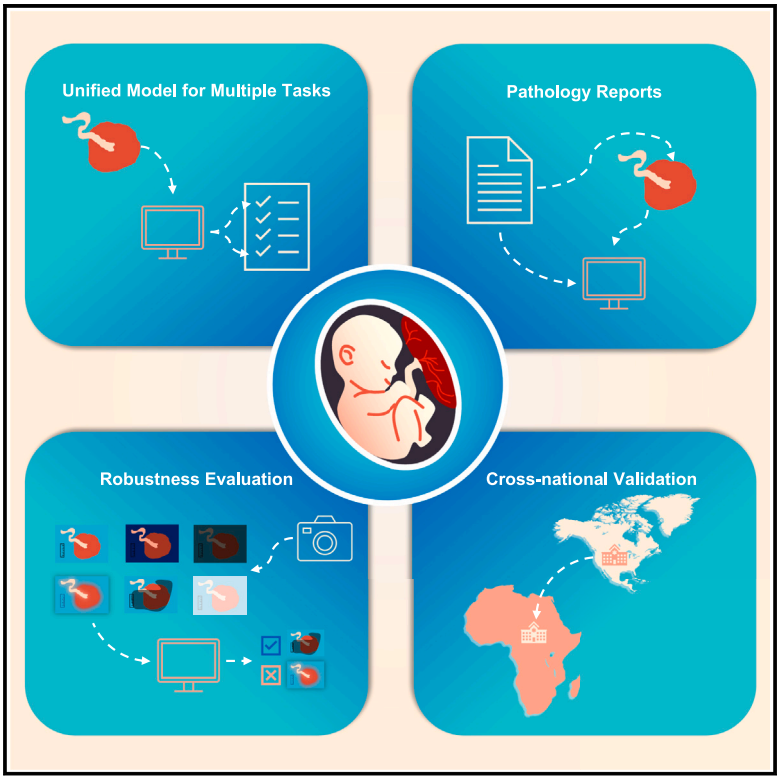
Article

Patterns

Cross-modal contrastive learning for unified placenta analysis using photographs

Graphical abstract



Highlights

- Unified placenta image encoder excels in placenta analysis
- A large, diverse dataset enhances the model's accuracy and robustness
- Robustness evaluation shows the model's resilience to various imaging conditions
- Cross-national validation confirms consistent performance across populations

Authors

Yimu Pan, Manas Mehta, Jeffery A. Goldstein, ..., Rachel E. Walker, Alison D. Gernand, James Z. Wang

Correspondence

adg14@psu.edu (A.D.G.), jwang@ist.psu.edu (J.Z.W.)

In brief

The placenta, vital for maternal and neonatal health, presents assessment challenges due to limited accessibility and expertise. This study presents a novel approach utilizing placenta photos spanning a 12-year period. Employing a unified model, it integrates cross-modal training between photos and pathology reports for diverse tasks. A proposed robustness evaluation protocol ensures method reliability. Validation across diverse populations underscores its potential for widespread clinical applicability.

Pan et al., 2024, Patterns *5*, 101097
December 13, 2024
https://doi.org/10.1016/j.patter.2024.101097

CellPress

Patterns

CellPress
OPEN ACCESS

Article

Cross-modal contrastive learning for unified placenta analysis using photographs

Yimu Pan,[1] Manas Mehta,[1] Jeffery A. Goldstein,[2] Joseph Ngonzi,[3] Lisa M. Bebell,[4,5] Drucilla J. Roberts,[5,6] Chrystalle Katte Carreon,[5,7] Kelly Gallagher,[8] Rachel E. Walker,[9] Alison D. Gernand,[9,*] and James Z. Wang[1,10,*]

[1]Data Sciences and Artificial Intelligence Section, College of Information Sciences and Technology, The Pennsylvania State University, University Park, PA, USA
[2]Department of Pathology, Feinberg School of Medicine, Northwestern University, Chicago, IL, USA
[3]Department of Obstetrics and Gynecology, Mbarara University of Science and Technology, Mbarara, Uganda
[4]Massachusetts General Hospital Department of Medicine, Division of Infectious Diseases, Medical Practice Evaluation Center, Center for Global Health, Boston, MA, USA
[5]Harvard Medical School, Boston, MA, USA
[6]Massachusetts General Hospital Department of Pathology, Boston, MA, USA
[7]Department of Pathology, Boston Children's Hospital and Harvard Medical School, Boston, MA, USA
[8]College of Nursing, The Pennsylvania State University, University Park, PA, USA
[9]Department of Nutritional Sciences, College of Health and Human Development, The Pennsylvania State University, University Park, PA, USA
[10]Lead contact
*Correspondence: adg14@psu.edu (A.D.G.), jwang@ist.psu.edu (J.Z.W.)
https://doi.org/10.1016/j.patter.2024.101097

THE BIGGER PICTURE The placenta plays a vital role in the health of both mother and baby during pregnancy, but it is often not thoroughly examined at birth, especially in resource-limited settings. This gap can lead to missed opportunities to detect critical conditions. Neonatal sepsis—a life-threatening infection—affects millions of newborns globally, particularly where early detection is challenging because of limited medical resources. This research introduces a powerful tool that enables quick and accessible placental assessment using just a photograph, potentially reducing the risk of undetected issues such as infection or placental abnormalities.

Adaptable into mobile applications, this innovation promises greater accessibility in both high- and low-resource settings. With further refinement, it has the potential to transform neonatal and maternal care by enabling early, personalized interventions that prevent severe health outcomes and improve the lives of mothers and infants worldwide.

SUMMARY

The placenta is vital to maternal and child health but often overlooked in pregnancy studies. Addressing the need for a more accessible and cost-effective method of placental assessment, our study introduces a computational tool designed for the analysis of placental photographs. Leveraging images and pathology reports collected from sites in the United States and Uganda over a 12-year period, we developed a cross-modal contrastive learning algorithm consisting of pre-alignment, distillation, and retrieval modules. Moreover, the proposed robustness evaluation protocol enables statistical assessment of performance improvements, provides deeper insight into the impact of different features on predictions, and offers practical guidance for its application in a variety of settings. Through extensive experimentation, our tool demonstrates an average area under the receiver operating characteristic curve score of over 82% in both internal and external validations, which underscores the potential of our tool to enhance clinical care across diverse environments.

INTRODUCTION

The placenta serves as a significant indicator of both pregnancy events and the health of the mother and baby.[1–13] However, even in a high-resource country like the United States, only approximately 20% of placentas undergo pathology examinations,[14,15] and placental data are often overlooked in pregnancy research.[16] The underutilization of placental pathology is primarily due to time, cost, expertise, and facility requirements, even in resource-abundant areas.[17] In low- and middle-income countries

(LMICs), the incidence of adverse maternal and newborn outcomes is higher, but resources are typically lacking to conduct placental pathology.[18,19] Therefore, enhancing the accessibility of placental assessment to pathologists, clinicians, and researchers is crucial.[20,21] Immediate placental assessment at birth is expected to significantly aid clinical decisions.

Existing automatic approaches in research often require expensive equipment and time (e.g., MRI,[22,23] computed tomography,[24,25] or histological images[21]) and are not suitable for immediate assessment after birth. Assessing abnormalities in the delivered placenta has substantial value by revealing events in pregnancy and labor that could impact clinical care for the postpartum mother and newborn. As an example, chorioamnionitis due to infection in the placenta may indicate a subclinical infection in the newborn. Histologic diagnosis of chorioamnionitis currently takes days[17] and therefore is not used to guide the immediate clinical care of the newborn. Due to the increased risk of early-onset neonatal sepsis,[26] antimicrobial agents are used even if the neonate appears well[27] before obtaining a diagnosis from placental pathology. A tool that could accurately estimate a diagnosis of chorioamnionitis (before a full pathology exam) and perhaps, more importantly, the absence of chorioamnionitis for the well-appearing newborn would help to initiate treatment for newborns most at risk of infection and appropriately limit antimicrobial use for those at low risk.

Recent efforts in placenta analysis have primarily focused on segmentation[28–30] and classification[31–37] using histopathological, ultrasound, or MRI images. Previous studies that utilized photographic images—a low-cost and immediate tool—to evaluate placental characteristics[38–41] and to perform placental diagnoses[42,43] required a separate model for each of these tasks due to the lack of a unified method. Moreover, the clinical outcomes that can be inferred from these placental diagnoses are often overlooked. A better utilization of the available data should explore the connection between the visual placental feature and the textual description from the pathology report independent of the downstream task. Additionally, using one model for multiple tasks would greatly save computational resources and improve the deployability of the resulting model. To fully leverage the information available in pathology reports and to train a unified placental feature encoder, our preceding work[44,45] introduced a vision-and-language contrastive learning (VLC) approach for placenta analysis. Similarly, in this work, we aim to further enhance the VLC approach in placental analysis toward robust deployment under various settings. VLC approaches[46,47] have garnered significant interest, particularly following the success of contrastive language and image pre-training (CLIP).[47] Research in this area has focused on improving VLC methodologies through innovations in model architectures,[48,49] visual representations,[50–52] textual representations,[45,53] and loss functions[54,55] as well as sampling strategies,[56,57] training strategies,[58] and classifier performance.[59] Another significant line of research[60–62] aims to improve the efficiency of VLC techniques. In the context of medical applications, a recent survey[63] provides a comprehensive overview of popular methods. Representative strategies have been proposed to enhance local feature alignment,[64–67] introduce auxiliary reconstruction tasks,[68,69] and incorporate external prior knowledge.[70–72]

Building upon our unified pre-training methods,[44,45] this research seeks to harness the simplicity and affordability of digital photography combined with the proposed cross-modal pre-training techniques to develop a unified model capable of comprehensive and immediate placental assessment using photographic images. Our contributions are as follows. (1) Introduction of a cross-modal contrastive learning technique designed to enhance both the performance and robustness of the placenta analysis model. (2) Development of three key modules: a cross-modal pre-alignment module for improved alignment between images and pathology reports using external image data, a cross-modal distillation module that leverages external textual information to capture nuanced relationships between placental features, and a cross-modal retrieval module for matching textual and visual features, fostering robust representation learning. (3) Creation of a robustness evaluation protocol tailored for placenta photographs to assess model robustness, facilitate explainability, and generate application guidelines. (4) Expansion of the training dataset 3-fold and broadening of the evaluation dataset to include a diverse, multinational collection, enhancing the model's applicability and performance in LMICs.

RESULTS

Dataset

The characteristics of the collected dataset are categorized in Figure 1. The "not applicable" (NA) category was excluded from the results. The primary dataset was collected in the pathology department at Northwestern Memorial Hospital (NMH) (Chicago, IL, USA) between January 1, 2010 and December 31, 2022, following the placenta imaging protocol.[73] Photographs were taken using a dedicated pathology specimen photography system (Macropath, Milestone Medical, Kalamazoo, MI, USA) with an integrated, fixed camera and built-in lighting to reduce technical variability. Pathologists generated the pathology reports based on histological findings and widely adopted definitions.[74] The refined dataset from NMH comprises 31,763 fetal-side placenta images, each accompanied by a pathology report. We selected 2,811 image-report pairs from the year 2017 for internal validation and the rest for pre-training.

Additionally, we identified 166 cases where the neonate was diagnosed with sepsis and 1,837 potential negative cases from the internal validation set. Furthermore, the external validation set was collected at the Mbarara Regional Referral Hospital (MRRH) (Mbarara, Uganda) between December 1, 2019 and November 30, 2023, under the Placentas, Antibodies, and Child Outcomes study using a Fujifilm FinePix XP130 digital camera. The imaging protocol is included in Data S1. The pathology reports were generated using the same method as described for the primary dataset. We obtained 353 placenta and pathology report pairs from MRRH for external validation. Following our preliminary research,[44,45] the AI-based placental assessment and examination (AI-PLAX) algorithm[42] was used to mask the background of each image in the NMH dataset. Additionally, the automatic and interactive segment anything model (AI-SAM) algorithm[75] was used to mask the background of each image in the MRRH dataset.

For the internal validation set, we manually checked the images to ensure that the placenta was complete and that its visibility was unobscured. We first labeled each image according to the

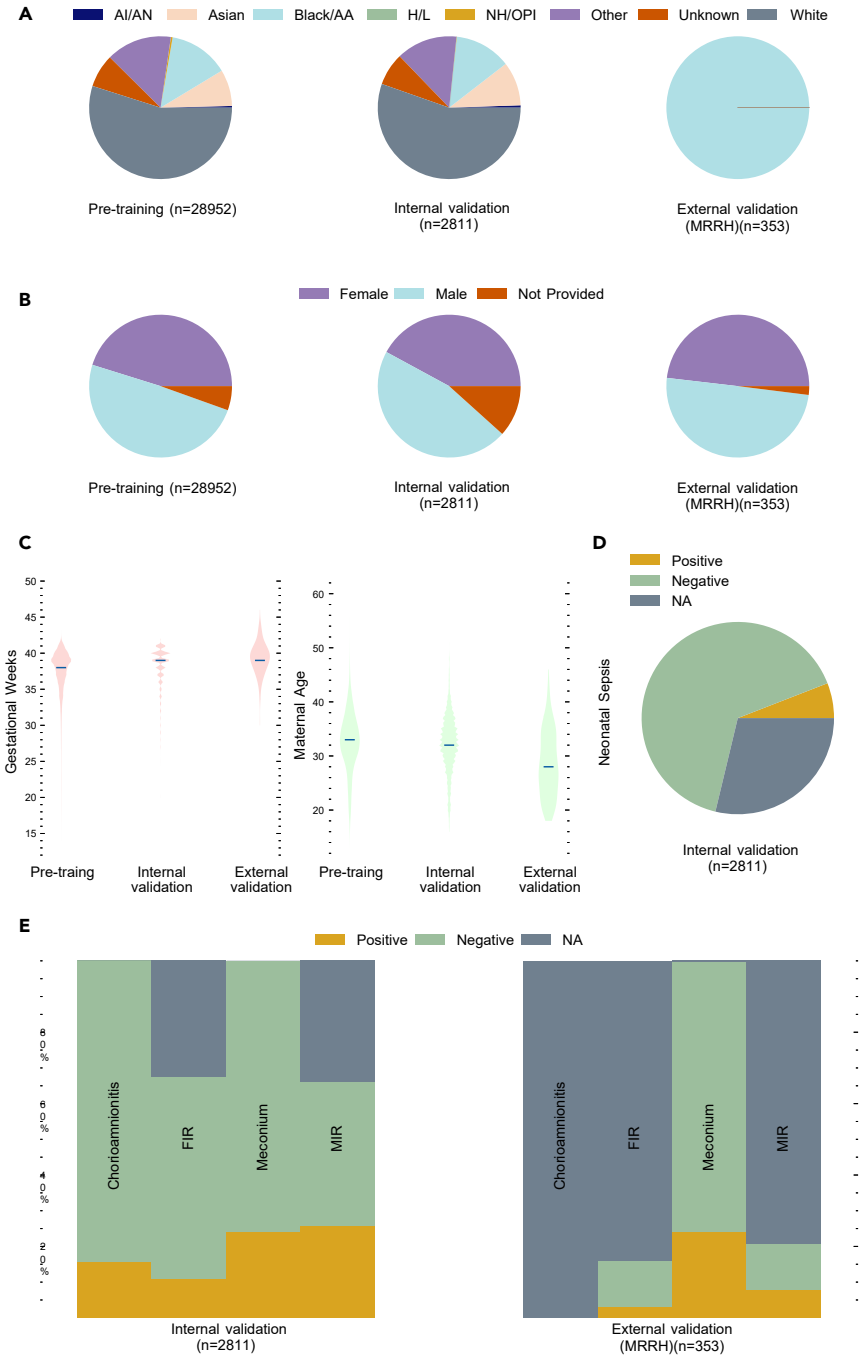


Figure 1. The characteristics of the primary dataset and the external validation dataset

(A) Distribution of self-reported race.

(B) Distribution of infant sex.

(C) Distribution of gestational age and maternal age.

(D) Distribution of the neonatal sepsis label from the tuning and validation set.

(E) Distribution of placental feature labels.

Each placenta from the primary dataset has one image and one pathology report, while placentas from the external validation dataset have a median (25%–75% percentile) of 4 (3–5) images and one pathology report. The NA category represents instances where information could not be derived due to missing data. AI/AN, American Indian or Alaska Native; Black/AA, Black or African American; H/L, Hispanic or Latino; NH/OPI, Native Hawaiian or other Pacific Islander; NMH, Northwestern Memorial Hospital; MRRH, Mbarara Regional Referral Hospital; FIR, fetal inflammatory response; MIR, maternal inflammatory response; NA, not applicable.

pathology report on four placental feature identification tasks outlined in previous work[42,43]; namely, meconium-laden macrophages in the amnion or chorion (hereafter referred to as meconium), fetal inflammatory response (FIR), maternal inflammatory response (MIR), and chorioamnionitis. Different levels or stages exist for some lesions. We labeled images as positive for meconium and chorioamnionitis regardless of the reported severity. For FIR and MIR, an image was labeled as negative if the report either lacked relevant information or explicitly indicated a negative diagnosis. Conversely, an image was labeled as positive if the report identified the placenta as stage 2 or higher. To enhance the model's ability to differentiate significant cases, images were excluded if their associated stage was 1.

A clinical outcome, neonatal sepsis, relates to placental features from the pathology report and is the most immediate cause of neonatal deaths in LMICs.[19,76] We retrieved images of cases with neonatal sepsis based on diagnoses made by treating physicians using clinical criteria from infant charts. We selected negative samples from the fine-tuning dataset that were free from FIR, MIR, and chorioamnionitis—the placental features related to sepsis and placental cause of death[77]—to minimize the possibility of having false negative samples. Evaluating the model's performance on such a task can infer its prediction capability on clinical outcomes that are related to placental features but not in the pathology report.

We used all positive examples for each task and uniformly sampled a comparable number of negative samples from the internal validation set to perform linear evaluation. The error range computation was based on random divisions of these data, with a 50:50 split for tuning and evaluating the linear classifier.

For the external validation set, we identified three placental feature identification tasks (namely, meconium, FIR, and MIR) and produced the classification labels using the same method. We excluded samples lacking corresponding pathology reports.

Model design

In general, we address two primary tasks: pre-training and downstream classification. Formally, in the pre-training phase, our objective is to learn a function f_v using another function f_u

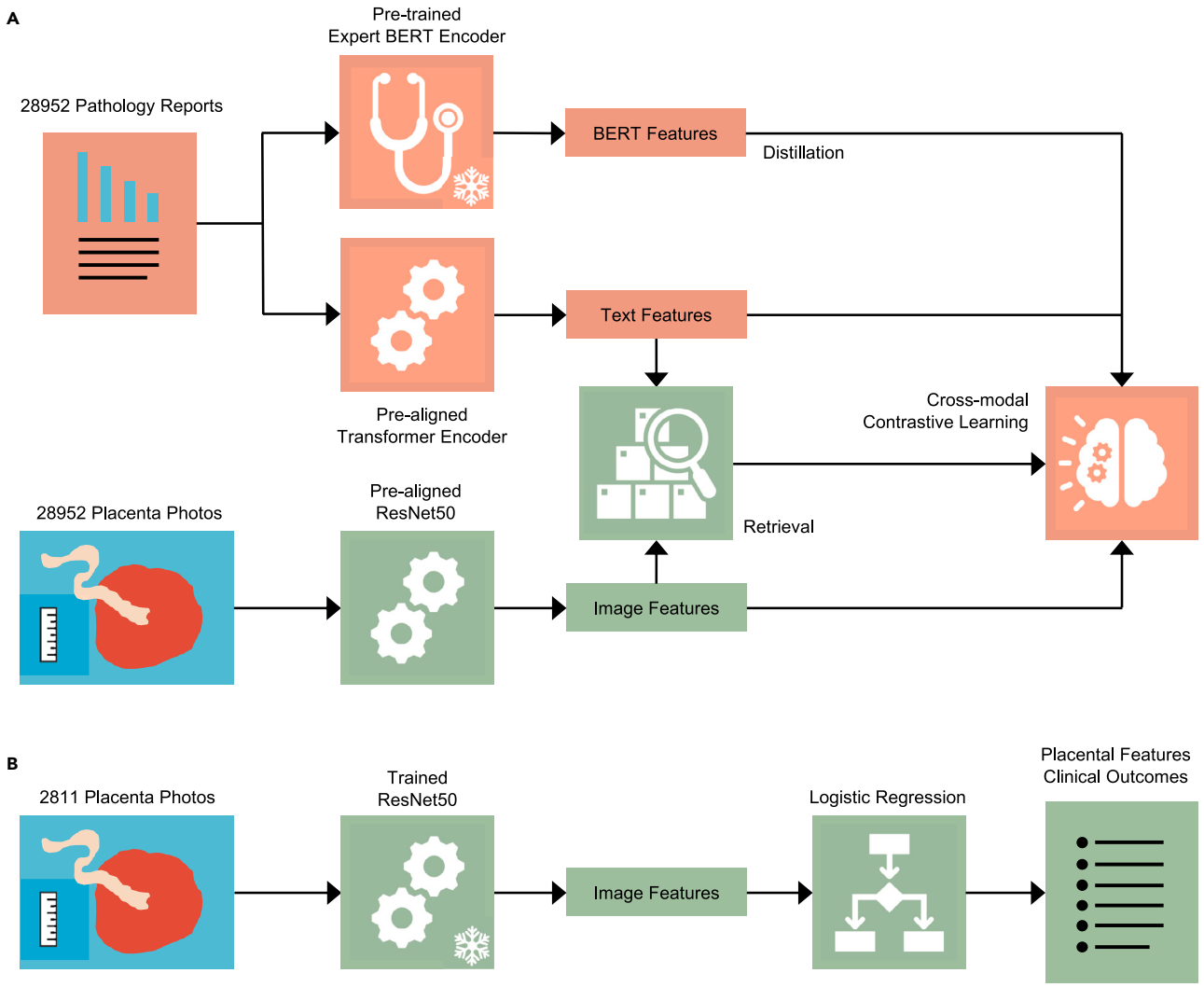


Figure 2. An overview of the pre-training and fine-tuning paradigm and the cross-modal contrastive learning algorithm PlacentaCLIP

(A) The pre-training stage, where a frozen pre-trained BERT encoder and a trainable transformer text encoder were used to encode the text from pathology reports, while a trainable ResNet50 was used to encode image features. The proposed cross-modal contrastive learning algorithm guides this training stage. BERT, bidirectional encoder representations from transformers.

(B) The fine-tuning stage, where the frozen ResNet50, trained in the pre-training stage, was used to extract image features, and logistic regression was applied to these features to predict the placental features or clinical outcomes. The 2,811-image fine-tuning dataset was randomly split into training and validation sets.

so that, for any given pair of inputs $(\boldsymbol{x}_i, \boldsymbol{t}_i)$ and a similarity function $\langle \cdot, \cdot \rangle$, we have $\boldsymbol{v}_i = g_v(f_v(\boldsymbol{x}_i))$ and $\boldsymbol{u}_i = g_u(f_u(\boldsymbol{t}_i))$. The objective function is defined as

$$\ell_i^{(v \to u)} = -\log \frac{\exp\left(\langle \boldsymbol{v}_i, \boldsymbol{u}_i \rangle / \tau\right)}{\sum_{k=1}^{N} \exp\left(\langle \boldsymbol{v}_i, \boldsymbol{u}_k \rangle / \tau\right)}, \qquad \text{(Equation 1)}$$

where τ is the temperature in the contrastive loss function, used to control the strength of contrastive learning.

In the downstream classification task, our goal is to learn a function f_{ct} using the learned function f_v for each task $t \in \{1, 2, \ldots, T\}$ so that, for a pair of input $(\boldsymbol{x}_i, l_{ti})$,

$$f_{ct}(f_v(\boldsymbol{x}_i)) = l_{ti}, \qquad \text{(Equation 2)}$$

which can be achieved by using a linear classifier.

Given that one of the objectives of placenta pathology reporting is to identify clinically significant findings and make diagnoses, an effective placenta photo analysis model needs to achieve comparable performance on both placental features identified in pathology reports and related clinical findings. To achieve these goals, we adopt the well-established pre-training and fine-tuning paradigm using contrastive learning techniques and propose the placenta feature encoder through CLIP (PlacentaCLIP).

During the pre-training stage (Figure 2A), we train a generalizable fetal-side placenta image encoder that is task agnostic. In the fine-tuning stage (Figure 2B), we train a simple classifier (logistic regression) using the encoded image features for each task. PlacentaCLIP builds on our preliminary work[44,45] by introducing cross-modal pre-alignment, distillation, and retrieval strategies. The cross-modal pre-alignment technique is designed to improve performance by pre-aligning the encoders, a

ResNet-based[78] image encoder, and a transformer-based[79] text encoder, with a large collection of natural image-text pairs.[47] This reduces the model's dependency on extensive placenta-specific data. The cross-modal distillation module distills intra-placental feature reasoning capabilities from a language model[80] trained on a large medical text corpus into the image encoder to enhance performance. Moreover, the cross-modal retrieval module improves robustness by retrieving the image regions relevant to the textual features for more effective image-text alignment. More details are provided in the methods section.

Robustness evaluation design

Placenta photographs are often subject to various non-ideal conditions that can affect their quality,[81] making it challenging for a modern analysis models to interpret them accurately. While a previous study[43] has tackled similar problems, our research offers a more comprehensive analysis. To evaluate the robustness of placenta analysis models in practical environments, it was necessary to account for these common variations and understand their impact on model performance. We developed a robustness evaluation dataset that includes common artifacts specific to placenta photographs, including motion blur, blood stains, and lighting variations. This evaluation provided valuable insights into the model's strengths and weaknesses, identifying areas for further enhancement. The evaluation protocol focused on three key objectives: first, to assess the robustness of the proposed modules (i.e., module evaluation); second, to identify potential correlations between tasks and image features (i.e., model explainability); and third, to offer guidance on optimizing photo-taking procedures in real-world applications (i.e., application guideline). The first two objectives contribute to model design and evaluation, while the third directly informs clinical practice. More details are provided under Robustness evaluation dataset generation.

To evaluate the model's performance under different artifacts, we intentionally introduced these variations or artifacts into the internal validation set. Unlike previous work,[82] we focused on creating a list of placenta-specific common corruptions based on our experiences and understanding of placenta photo-taking procedures, aiming to more accurately simulate real-world settings. We divided the corruptions into three groups: image artifacts (blood, glare, JPEG compression, and shadow), image blur (defocus, motion, and zoom), and exposure artifacts (brightness, contrast, and saturation). Each corruption was assigned five levels, with each level representing a different degree of severity. It was important to set realistic corruption strengths to ensure that our evaluation of the model's robustness was relevant to practical placental analysis. First, we chose levels that maintained placental visibility across all corrupted images. Then, through consultation with a pathologist, we determined the highest corruption level at which placental features remained discernible. These levels were standardized to level 3 (of 5), and the other levels were adjusted accordingly. Additionally, we included common white balance inaccuracies (e.g., tungsten, fluorescent, daylight, cloudy, and shade), simulated using white balance augmentation.[83] Further details and examples are provided in the methods section.

Model performance and robustness evaluation

The performance of the linear classifiers was compared and quantified using the area under the receiver operating characteristic curve (AUC), mean average precision (mAP), and 1 − Brier score[84] to ensure that the threshold for positive predictions did not affect the scores. We measured robustness by observing the performance drop when introducing image artifacts.

Overall performance

PlacentaCLIP was trained using 10,193 image-text pairs from 2014 to 2018, in alignment with previous work.[44,45] PlacentaCLIP+ was trained on 28,952 image-text pairs from 2010 to 2022 to demonstrate its full capability and scalability. The results in Figure 3A indicate that PlacentaCLIP achieves state-of-the-art AUC. Additionally, the AUC improvements from incorporating additional data (PlacentaCLIP+) highlight the scalability of our proposed method. Figure 3B illustrates how the cross-modal retrieval module aids the pre-training process. We visualize the attention weights from each text query to the image features to demonstrate the changes in the feature space. Including stage 1 for FIR and MIR increased the variance of the model performance, but the performance was similar. This result is provided in Table S1.

Application guideline from robustness evaluation

As shown in Figure 4, the introduction of any form of corruption adversely affected the model's performance. Despite significant corruption, factors such as shadows, zoom blur, contrast, and saturation changes exerted a relatively minor impact on performance. These results offered new insights into the photo-taking process, challenging general assumptions. Notably, commonly used lossy image compression techniques (e.g., JPEG) significantly degraded the model's performance, even though the input size (512 × 384) of the model was smaller than images produced by most smartphones. Consequently, users are advised to avoid lossy compression and instead opt for smaller image sizes if storage space is a concern. Additionally, glare had a more detrimental effect on the model's performance than shadows, suggesting that shielding light sources to reduce glare could improve performance. The model also showed greater sensitivity to brightness changes compared to contrast and saturation, implying that adjusting contrast and saturation could be a viable method to compensate for brightness issues. Each aspect of robustness was assessed in isolation to avoid complex factorial comparisons and to facilitate the generation of application guidelines. Initial analysis on the combined aspects are included in Tables S2–S6, and original statistical results are shown in Table S8; however, no additional insights were gained due to the complexity of analyzing the large number of comparisons.

As shown in Figure 4D, the model's performance degradation under white balance inaccuracies was moderate and consistent, except when the white balance preset used was extremely inaccurate (e.g., tungsten 2850K). Moreover, when the white balance preset aligned closely with the actual lighting conditions (e.g., between fluorescent and daylight), the model demonstrated its best performance. This indicated that the model possesses a degree of adaptability to various white balance inaccuracies. Nevertheless, users are advised to verify the white balance setting for optimal performance or use the camera's automatic setting to secure reasonable performance.

Model explainability from robustness evaluation

Figure 4 illustrates that each task exhibits distinct levels of robustness against various artifact. Sepsis prediction demonstrated the

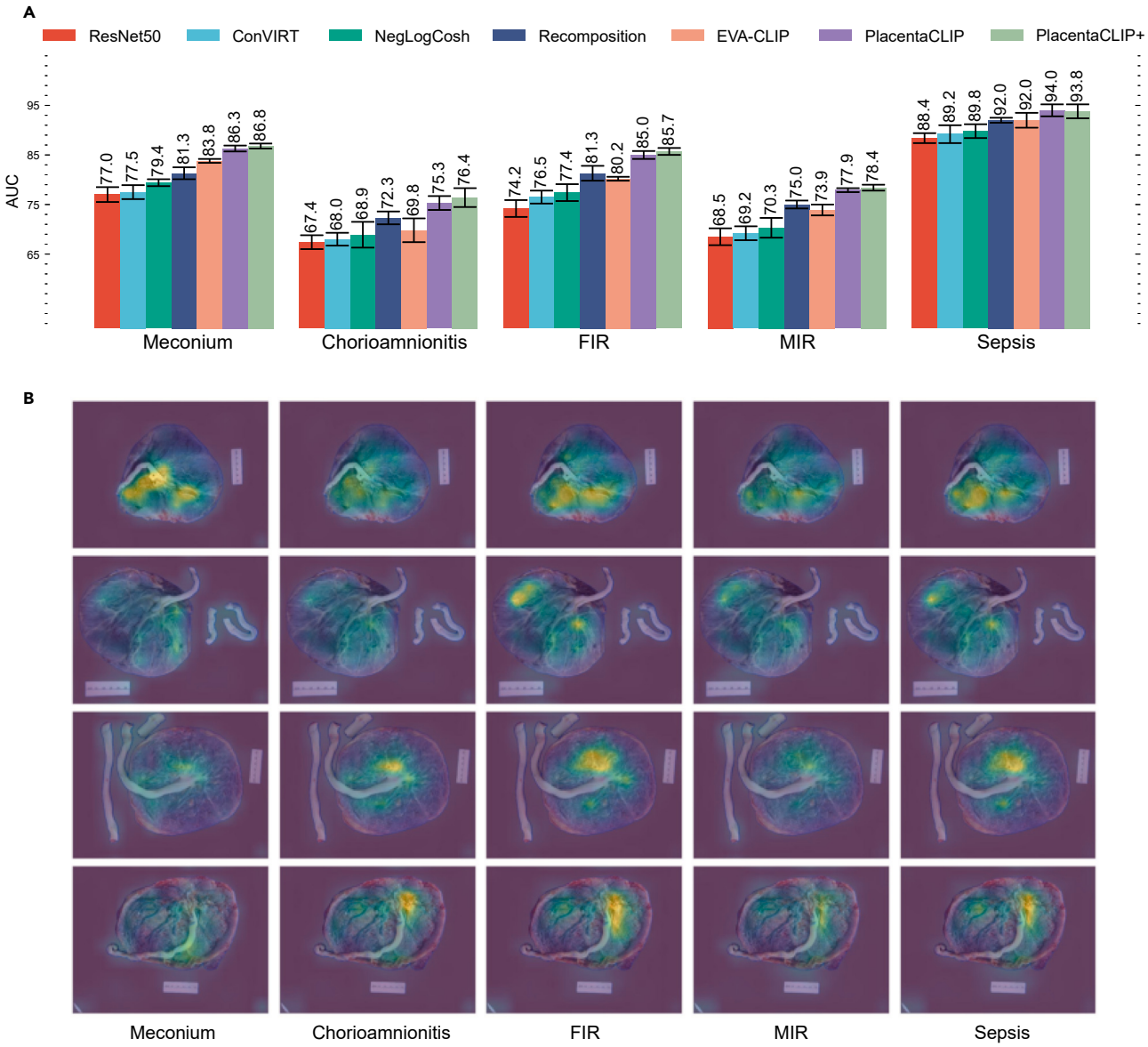


Figure 3. Average AUC for four placental feature identification tasks and one clinical outcome prediction task on the primary dataset and visualization of the cross-modal retrieval module

(A) The AUC and the corresponding standard deviation from five random splits. Results of ResNet50, ConVIRT,[46] and NegLogCosh are taken from Pan et al.[44] Results of recomposition are taken from Pan et al.[45] The result for EVA-CLIP[58] is from the "EVA02_CLIP_B_psz16_s8B" model, tuned on our pre-training data. The error bars represent the standard deviations computed from five random splits.

(B) The attention weights from the cross-modal retrieval module during the pre-training stage. Different features are retrieved to assist the image encoder pre-training based on query text for better image-text alignment. In the illustration, the full name of each task is used as the text query to retrieve the visual features, except for sepsis, where the concatenation of FIR, MIR, and chorioamnionitis is used as the query. The actual process uses part of the report as the text query. The ground-truth labels for the images from top to bottom are as follows: row 1: –, –, –, –, –; row 2: +, –, –, –, –; row 3: –, –, 1, 1, –; row 4: +, –, 1, 1, –. –: negative; +: positive; 1: stage 1. FIR, fetal inflammatory response; MIR, maternal inflammatory response; PlacentaCLIP+, PlacentaCLIP trained with additional data.

highest overall robustness, likely because the model depends on multiple placental features to predict sepsis risk, which weakens the effect of individual artifacts. Meconium was particularly sensitive to color alterations (e.g., saturation changes in Figure 4C and extreme white balance inaccuracies in Figure 4D), whereas MIR was more susceptible to the loss of textural details due to JPEG compression, as shown in Figure 4A. Moreover, blur significantly compromised performance by removing textural details. As depicted in Figure 4B, tasks like FIR, MIR, and chorioamnionitis were less robust to blur compared to meconium, underscoring their reliance on textural details. Zoom blur, which affected images non-uniformly, impaired textural features in the outer part

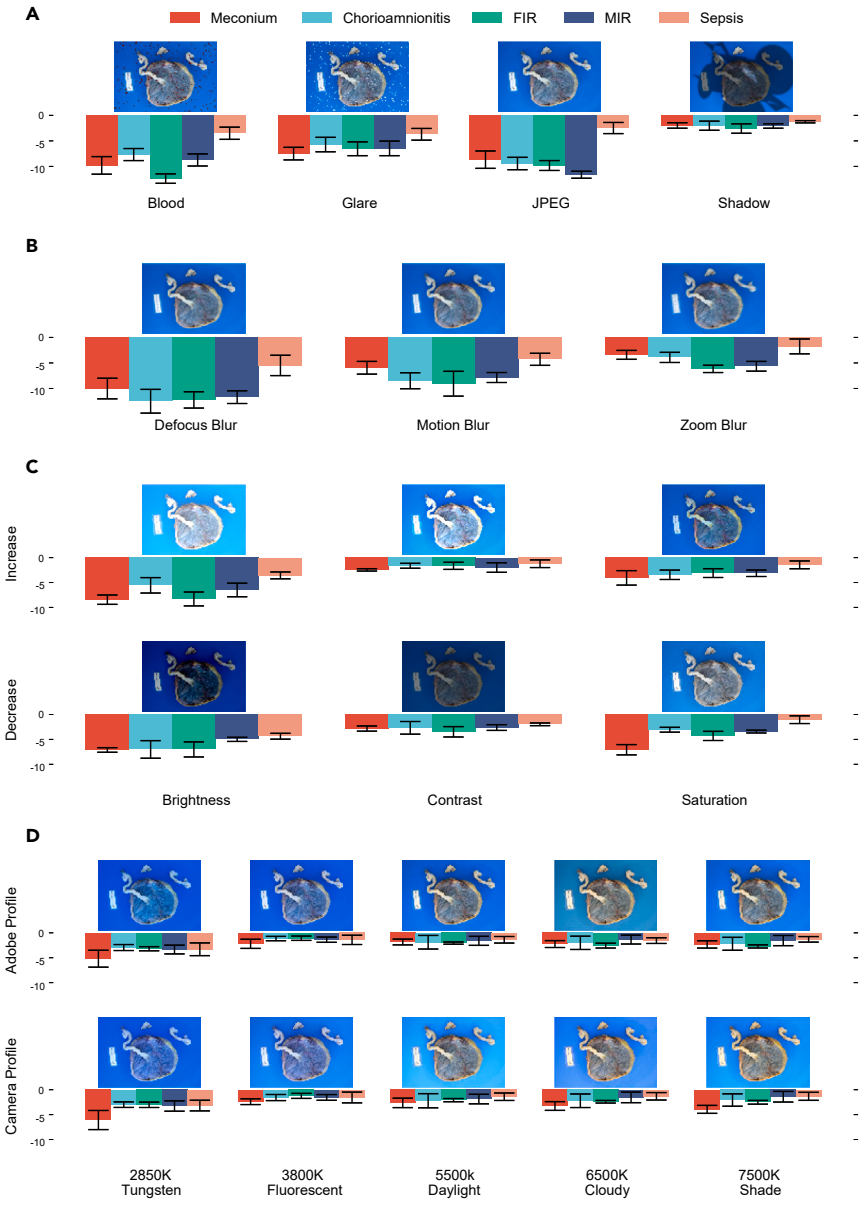


Figure 4. The average AUC performance drop of PlacentaCLIP+ from using the original images to corrupted images on each task identified in the primary dataset

The AUC drop (y axis) is computed by subtracting the AUC of PlacentaCLIP+ on the original images from the AUC on corrupted images, averaged across all corruption levels for each random split. Error bars represent the standard deviations computed using five random splits.

(A) Performance under different image artifacts.

(B) Performance under different types of image blurring.

(C) Performance under different exposure artifacts.

(D) Performance under different WB inaccuracies.

FIR, fetal inflammatory response; MIR, maternal inflammatory response.

of the image, particularly where the umbilical cord is present, more than the center. The lesser impact of zoom blur compared to uniform blur suggests a stronger reliance on the features in the center placenta disk than on the cord for all tasks. Notably, while the relative performance between other tasks stayed consistent across all three types of blur, chorioamnionitis exhibited a much lower performance drop with zoom blur, suggesting greater robustness to blur in the cord region. This finding was further supported by the relative performance change between tasks under varying brightness levels. Increasing brightness (or overexposure) removed information more quickly from brighter regions, while decreasing brightness (or underexposure) similarly affected darker regions. Chorioamnionitis was more adversely affected by decreased brightness than increased brightness, suggesting that the model relies more on darker regions (e.g., the disk) than the brighter regions (e.g., the cord) for predicting this condition. This result partially aligns with pathological examination, where MIR and chorioamnionitis are found in the disk region, while FIR at stage 2 or higher is primarily found in the cord.

Module contribution to performance and robustness

To further investigate the differential behavior of the proposed modules, we analyzed their performance across varying corruption levels as part of the robustness evaluation protocol. Figure 5 demonstrates how each module responded to changes in corruption severity. Specifically, applying the cross-modal distillation module (PD) yielded superior AUC scores at lower corruption levels (levels 0–2) compared to the retrieval module (PR). Conversely, at higher corruption levels (levels 3–5), the retrieval module outperformed the distillation module in terms of AUC. This variation in performance suggests that the retrieval module primarily enhances robustness, while the distillation module boosts performance. Importantly, the two modules are complementary, as their combined use (PDR) resulted in higher AUC scores. Finally, these modules demonstrated scalability, with the inclusion of additional data (PDR+) improving both performance and robustness.

Module hyperparameter evaluation

To analyze the effects of the hyperparameter for each proposed module, we trained the model with each module individually across a range of parameter settings, as shown in Figure 6. The x axis represents the hyperparameter values regulating the strength of each loss function, while the y axis depicts the corresponding model performance. The PDR line represents the performance of the final model with all modules and additional data incorporated, and the P line indicates the performance of the baseline pre-aligned model without any proposed modules. It is observed that the performance of the model with individual

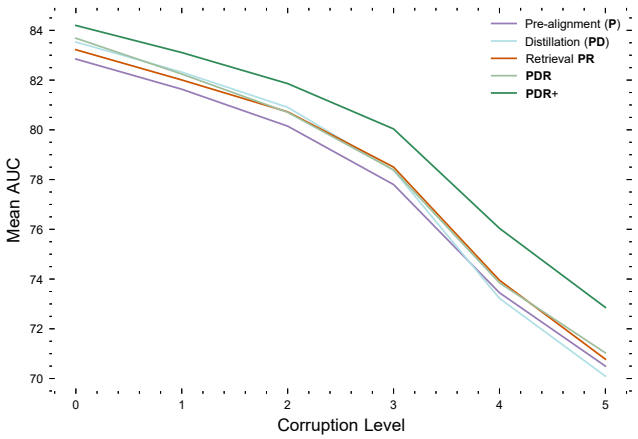


Figure 5. Module AUC performance at varying corruption levels

At lower corruption levels (below level 3), the distillation module outperformed the retrieval module. As the corruption level increased, the retrieval module showed better performance. Adding the distillation module on top of the retrieval module did not further improve robustness (i.e., the performance of PDR and PR converged as the corruption level increased). These results validate our design; distillation enhances performance, while retrieval improves robustness.

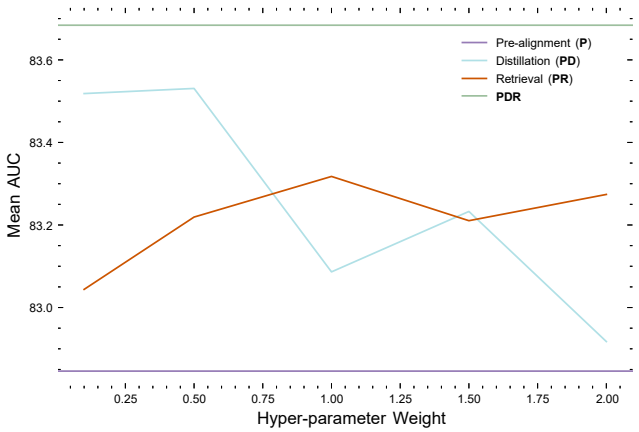


Figure 6. The AUC scores obtained by applying different hyperparameters to the pre-training modules on a subset of the primary dataset

modules falls between these two lines for all parameter settings experimented.

Particularly noteworthy is the trend observed in the distillation line. We noted a general decline in performance as the weight of the loss increased. This aligns with our expectations; introducing a distillation loss from a pre-trained model enhances reasoning capabilities, but an excessively high weight assigned to this loss can lead to a considerable domain shift from the pre-trained language model. This shift can adversely affect model performance. If the weight for the distillation loss is set too high, then it effectively reduces to a distillation-only loss, as seen in our previous work.[44] Thus, it is preferable to assign a lower weight for this module. In contrast, the cross-modal retrieval module exhibits a different pattern; model performance improves and then plateaus as the weight of this loss increases. This trend, opposite to that of distillation, is anticipated, since this module is linked to trainable parameters and does not introduce domain shift. Increasing this loss introduces additional priors, but beyond a certain point, it adds no further benefit. Consequently, further performance improvement are not anticipated after a certain threshold, making it advisable to select a more conservative weight for this module.

Statistical analysis

We contrasted each module's performance on the primary dataset and under the robustness evaluation protocol to assess whether they met their intended design rationales. Furthermore, to identify potential biases within the model, we examined its performance in relation to the demographic information present in the primary dataset. We used a paired t test with a significance level of 0.05 applied to the AUC scores. This involved treating the five classifiers trained on different splits as individual subjects. The t test was conducted as a within-subjects test, considering that all other configurations remained identical except for the test variables. Additionally, we applied the Benjamini-Hochberg procedure[85] to all multiple tests to control the false discovery rate. All results are presented in the following sections.

Module contribution to performance and robustness

Performance was measured using the internal validation set, while robustness was assessed using the robustness evaluation set (i.e., internal evaluation with introduced artifacts). The use of the cross-modal retrieval module (R) significantly strengthened robustness, as indicated in rows 1 and 2 of Table 1, though its impact on performance enhancement was comparatively modest. Conversely, rows 2 and 3 demonstrate that the use of the distillation module (D) led to a significant uplift in performance but had a limited affect on robustness. These results are consistent with our model designs. Additionally, the enhancements in accuracy and robustness provided by the proposed modules were complementary (DR). When both modules were applied (row 6), there was a significant improvement in both accuracy and robustness. Last, the proposed modules demonstrated scalability, as the addition of extra data (+) into the training set consistently resulted in significant gains in both accuracy and robustness (rows 7–10). These findings align with the robustness evaluation results from Module contribution to performance and robustness.

Bias assessment

In the bias assessment, our aim was to identify any potential biases in model performance across different demographic groups. Statistical analysis was conducted on the main dataset for various demographic categories.

We first analyzed the mean AUC across all tasks, considering race as the within-subject factor and each model trained on a different random split as the subject. The t test results, presented in Table 2, indicated no statistically significant differences in model performance among the known racial groups. Only the performance for the "unknown" racial category was notably better than that for the "White" group. Additional analysis (Table S7) revealed that the disparity in class distribution for the sepsis classification task contributed to this performance difference.

Next, the mean AUC across all tasks was compared using maternal age groups (in years) as defined in the obstetric care consensus,[86] with each model trained on a different random split

Table 1. Pairwise t test results applied to the AUC between applying set A of modules and set B of modules on both the primary dataset and the robustness evaluation

				Performance		Robustness	
	A	B	DoF	T score	*p*-Corr	T score	*p*-Corr
1	PD$\overline{R}$	PD	4.0	1.351	0.248	3.880	0.030
2	P$\overline{R}$	P	4.0	2.874	0.057	3.452	0.037
3	P$\overline{D}$R	PR	4.0	4.010	0.023	0.970	0.385
4	P$\overline{D}$	P	4.0	4.569	0.017	1.435	0.250
5	P$\overline{D}$	P$\overline{R}$	4.0	2.718	0.059	−1.811	0.181
6	P$\overline{DR}$	P	4.0	20.550	0.000	4.031	0.030
7	P$\overline{DR}$+	P	4.0	16.077	0.000	12.114	0.001
8	PD$\overline{R}$+	PD	4.0	5.631	0.010	9.442	0.002
9	P$\overline{D}$R+	PR	4.0	7.930	0.003	13.286	0.001
10	PDR+	PDR	4.0	8.237	0.003	11.273	0.001

The performance of all modules is reported using the same image encoder. Cross-modal retrieval improved robustness, while cross-modal distillation enhanced the performance of the image encoder. Performance was measured using the average AUC score for each set of modules on the original primary dataset. Robustness was measured using the average AUC score for each set of modules on the corrupted primary dataset (robustness evaluation protocol). P represents the use of pre-aligned encoders, D indicates cross-modal distillation, and R refers to cross-modal retrieval. PDR+ denotes the use of additional data in conjunction with all modules. A bar is placed above the differing module. DoF, degree of freedom; *p*-corr, corrected *p* value.

Table 2. Pairwise t test comparing the AUC between demographic groups A and B

A	B	T	DoF	*p*-Corr
Black/African American	American Indian/ Alaska Native	0.885	4.0	0.502
Black/African American	White	0.832	4.0	0.502
White	American Indian/ Alaska Native	0.737	4.0	0.502
Asian	American Indian/ Alaska Native	1.329	4.0	0.358
Asian	Black/African American	1.735	4.0	0.338
Asian	White	3.387	4.0	0.103
Other	American Indian/ Alaska Native	1.303	4.0	0.358
Other	Asian	0.754	4.0	0.502
Other	Black/African American	2.037	4.0	0.278
Other	White	4.049	4.0	0.077
Unknown	American Indian/ Alaska Native	1.588	4.0	0.351
Unknown	Asian	2.737	4.0	0.156
Unknown	Black/African American	4.737	4.0	0.068
Unknown	other	1.391	4.0	0.358
Unknown	White	9.690	4.0	0.010

Performance was measured using the average AUC score for each set of modules on the original primary dataset. DoF, degree of freedom; *p*-Corr, corrected *p* value.

as the subject. As shown in Table 3, the model demonstrated improved performance in the age group of 45 and above. There was no clear biological explanation for this. However, the increase in average age from pre-training to the internal validation set, as shown in Figure 1C, may have contributed to this performance difference.

Additionally, the analysis involved comparing the mean AUC across all tasks with gestational age groups (in weeks), as outlined in the committee opinion,[87] as the within-subject factor. The results revealed no statistically significant differences in model performance across the different gestational age groups.

Finally, we assessed the mean AUC across all tasks, using fetal sex as the within-subject factor, with each model trained on a different random split as the subject. The analysis yielded a t score of 1.069 and a p value of 0.345, indicating no statistically significant difference in model performance based on fetal sex.

External validation

To evaluate the model's performance in practical application scenarios, where users may capture placental images under diverse conditions, we conducted tests using the dataset from MRRH. The distinctiveness of this dataset lies in its provision of multiple images, varying in quality, for each placenta. For our analysis, we computed the best, worst, and mean performance metrics. The best and worst performances were determined by selecting the images where the model achieved the highest and lowest effectiveness for each placenta, respectively. Meanwhile, the mean performance was calculated by averaging the predicted probabilities across all images for each placenta.

As shown in Figure 7A, the mean performance on the MRRH dataset, which we regard as the most representative metric, was satisfactory but lower than expected. Apart from the domain shift from the training data, this reduced performance is attributable to variations in image quality, as evidenced by the substantial gap between the best and worst results and the variations in Figure 7B. Thus, it is reasonable to anticipate that the model's real-world performance in various clinical settings, on good-quality images, would fall between the mean and the best-observed results.

DISCUSSION

This study advances placenta analysis by introducing new data, models, and evaluation techniques. The integration results in a unified placenta analysis model with promising capabilities for placental feature identification and neonatal sepsis prediction, validated through our robustness evaluation protocol and a cross-national dataset. Unlike previous methods,[47,64–69] our approach synchronizes both internal and external representations. Moreover, our method extracts external knowledge from pre-trained language models without human intervention, whereas other methods[70–72] often rely on expert input. This distinction improves the generalizability of our approach, particularly in scenarios where expert knowledge is scarce or difficult to obtain. Moreover, this study examines robustness across

Table 3. Pairwise t test comparing the AUC of PlacentaCLIP+ using images from maternal age groups A and B

A	B	T	DoF	*p*-Corr
45 ≤ MA	35 ≤ MA < 40	6.936	4.0	0.007
45 ≤ MA	40 ≤ MA < 45	4.460	4.0	0.022
45 ≤ MA	MA < 35	7.205	4.0	0.007
35 ≤ MA < 40	40 ≤ MA < 45	1.465	4.0	0.217
MA < 35	35 ≤ MA < 40	2.417	4.0	0.110
MA < 35	40 ≤ MA < 45	2.126	4.0	0.121

Performance was measured using the average AUC score for each set of modules on the original primary dataset. MA, maternal age; DoF, degree of freedom; *p*-Corr, corrected *p* value.

different application settings, a critical concern in medical imaging. Our robustness analysis clarifies the factors that influence performance, providing insights for future model design. Finally, our findings offer guidance for clinical photography, identifying significant factors such as glare, JPEG compression, and blur, while noting lesser impacts from contrast, shadow, and minor white balance variations.

Clinical implementation and global health implications

In the United States and most birth settings around the world, the placenta undergoes only a brief visual examination after delivery.[14,15] Clinicians often receive minimal training on what to look for in the placenta,[17] focusing primarily on obvious signs such as incomplete sections that may indicate retained placenta. Generally, only a small proportion of placentas—around 20% in the United States—are sent for a full pathological examination, which takes 2–4 days to complete[14,15]; the remainder are discarded. In low-resource settings, such as Uganda, pathology departments may lack the capacity to examine placentas entirely, or such examinations are performed rarely, potentially missing crucial information about the pregnancy that could influence health outcomes.[26,27] Another issue is that hospital protocols for identifying which placentas should undergo pathological examination are often ineffective, with clinicians either unaware of College of American Pathologists guidelines or their own institutional guidelines, leading to the selection of placentas that do not provide the most critical information.[20] This results in wasted resources and missed opportunities to examine placentas with significant clinical relevance.

We aim to further refine the PlacentaCLIP+ algorithm to eventually integrate it into a mobile app that clinicians worldwide could use at the bedside for real-time, clinically relevant diagnoses concerning the placenta and maternal and neonatal health immediately after birth.[42,44,45] This would augment the clinician's placental expertise, allowing the model to identify important abnormalities, such as incomplete placentas or signs of infection.[42] Due to its ease of use, the app could be beneficial in any delivery setting worldwide, with the potential to significantly reduce morbidity and mortality. For example, by enabling the early identification of undetected incomplete placenta or infection risks, providers could intervene more quickly to reduce hemorrhage and sepsis rates. In high-resource settings, PlacentaCLIP+ could be used to triage placentas for full histopathological examination. In this workflow, obstetric providers would photograph the placenta in the delivery room and use the findings, along with their clinical judgment, to determine whether to submit the placenta for further examination. This process could increase the proportion of clinically relevant placentas sent to pathology and enable prioritization for rapid examination.

If we can sufficiently refine PlacentaCLIP+ to provide strong predictive value within an app, then we would evaluate its feasibility with the target end-users—clinicians working across various settings. Its similarity to existing smartphone applications that use photographs, such as those for submitting bank checks or receipts, should ease adoption and simplify training (i.e., minimal training would be needed for clinicians to take a photograph of the placenta). Point-of-care use would be straightforward, and integration with electronic health records could be achieved through the Health Level 7 system.[88] While introducing the app might add to the birth workflow, we anticipate that the tool and process could be streamlined to take minimal time (approximately 5 min) due to its simplicity and similarity to existing smartphone applications. Ultimately, it could save time and resources by reducing the incidence of poor health outcomes at birth.

PlacentaCLIP+ could also be valuable in research on pregnancy, birth, and childhood, providing a cost-effective way to include placental pathology and improve our understanding of long-term maternal and child health outcomes. A substantial body of evidence demonstrates links between placental pathologies/features and pregnancy complications,[8–11,89,90] risk of recurrence,[12,91] risks to the future health of the mother[13,92] and child,[93–95] and adverse long-term offspring health outcomes.[7,96–102]

Privacy and security concerns related to PlacentaCLIP+ would be minimal, as the algorithm analyzes images of a discarded organ. However, if clinical or demographic data were incorporated into the app, then additional measures would be required to ensure patient privacy and data security. PlacentaCLIP+ is part of our broader PlacentaVision project[103] aimed at enhancing the timely diagnosis of maternal, placental, and neonatal conditions that could affect health outcomes. Our findings suggest that bias across different ethnic and racial groups presents a relatively low risk. We are actively expanding our datasets from various sites to improve the algorithm's generalizability. PlacentaCLIP+ likely meets the definition of software as a medical device[104] by the US Food and Drug Administration (FDA). If new FDA regulations are implemented, then clearance would be required before its use in the United States. However, given our focus on global health, we anticipate that the algorithm will be of significant value worldwide before we pursue FDA clearance.

Limitations of the study

Despite its advancements, the model still exhibits a performance decline under the robustness evaluation protocol, suggesting potential areas for further enhancement. Ideally, the model's performance should mirror the consistency of a pathologist's ability to interpret visual cues under varying conditions. This gap in robustness compared to human experts may be attributed to biases in the training data and limitations in the model architecture, where images are of uniform quality and captured under consistent lighting conditions. To address this, increasing the

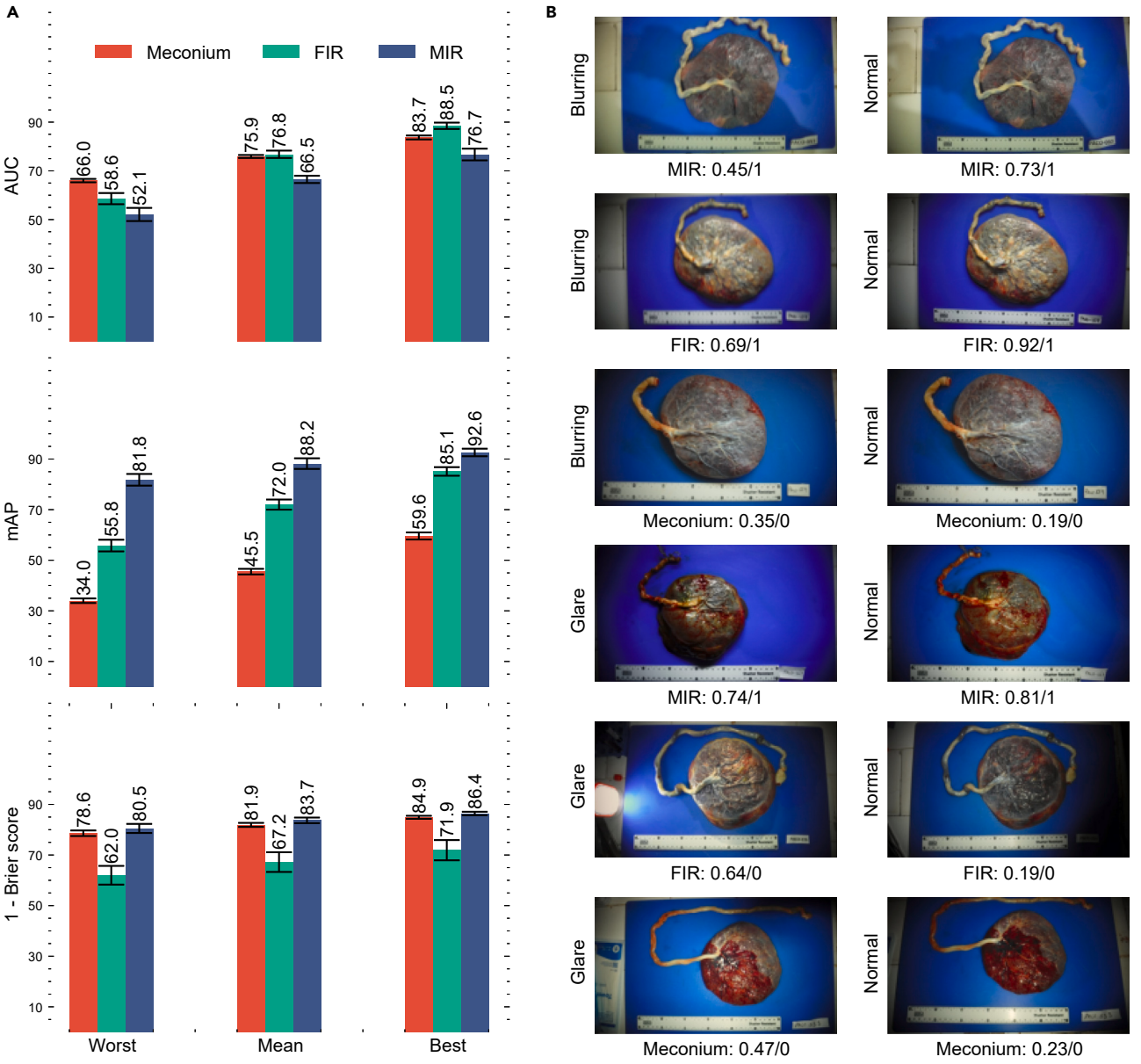


Figure 7. Performance of PlacentaCLIP+ on three placental feature identification tasks in the external validation set from MRRH and some qualitative examples of performance variation

(A) The performance for the three identified tasks using the models trained on the NMH dataset. Worst: the metrics were generated by selecting an image for each case where PlacentaCLIP+ performed the worst. Mean: the metrics were generated by averaging the probabilities predicted by PlacentaCLIP+ over all the images for each case. Best: the metrics were generated by selecting the image for each case there PlacentaCLIP+ performed the best. AUC, area under the receiver operating characteristic curve; mAP, mean average precision.

(B) Example performance and image quality variation. The examples on the left are more affected by the identified artifacts than those on the right. The reported model performance under each image is presented in the form of a task: prediction/ground truth.

diversity of the training dataset or introducing additional regularization into the model architecture could potentially enhance its robustness.

While the proposed robustness evaluation protocol is comprehensive, it inevitably has limited scope due to practical constraints and the complexity of real-world scenarios. In the future, a more dynamic and adaptive protocol, similar in concept to Autoaugment,[105] could be developed to generate meaningful combinations of aspects for a more thorough evaluation and explanation of a model's capabilities. Additionally, while the external validation dataset, which comprises images of varying quality, offers valuable insights into real-world model performance, it lacks sufficient diversity across demographic groups and devices. Another limitation of our study is its exclusive focus on the fetal side of the placenta without incorporating additional clinical data and considering the maternal side. Given the

established correlations between clinical data and placental outcomes[106] as well as the relevance of maternal-side placental features to health outcomes,[77] expanding the analysis to both sides of the placenta and incorporating clinical data is a critical next step.

Conclusion

In conclusion, this study presents a comprehensive analysis and three enhancements to a placenta analysis model. With proper photo-taking techniques,[73] the model's ability to accurately analyze placental images captured under various conditions, as indicated by the robustness evaluation protocol and the external validation results, makes it particularly suitable for environments where access to high-quality imaging equipment and expert medical personnel is limited. By leveraging commonly available devices like smartphones and tablets for image capture, this model can bridge the gap in placental assessment in regions where traditional pathology resources are scarce or nonexistent. The potential of this model to improve neonatal care in low-resource environments, where such advancements are most urgently needed, is particularly promising. This holds significant implications for advancing equitable and accessible maternal-fetal healthcare on a global scale with the potential to transform neonatal care in LMICs.

METHODS

Ethics statement

This work was conducted under Penn State single institutional review board (IRB) approval (STUDY00020697). The primary data collection was conducted under Northwestern IRB approval (STU00207700 and STU00215628). The external validation data collection was conducted under MUST REC MUREC 1/7 and Mass General Brigham IRB (2019P003248).

Module design and motivation

Cross-modal pre-alignment using natural images

In our previous method,[44] we trained the image encoder (f_v) using an unmodifiable ("frozen") pre-trained text encoder (f_u), similar to the ConVIRT method.[46] To enhance the robustness and generalizability of the model, we incorporated the NegLogCosh similarity and used sub-features. This approach was viewed as a form of distilling knowledge from the text encoder to the image encoder. However, in the context of placenta analysis, a frozen text encoder pre-trained on other tasks may not adapt to the specific demands of this domain without fine-tuning. A frozen text encoder does not learn the information specific to placenta features during training, which can limit both the accuracy and generalizability of the model. Moreover, pathology reports are highly structured, which may not suffice for training a language model without compromising its reasoning capability. Therefore, directly aligning trained text features with untrained image features may not be reasonable, as the text and image encoders may not be calibrated to a common conceptual space. Overcoming this misalignment would require a large corpus of image-text paired training data, which is prohibitively costly for placenta images and pathology reports. To navigate these constraints, in our work, we used a cost-effective dataset of natural image-text pairs sourced from the internet as the initial training data to pre-align the encoders and used our placenta data to shift the encoders to our specific placenta domains. To conserve computational resources, we started with the CLIP[47] ResNet50 and transformer models, which have already been trained on 400 million image-text pairs. Through continued training of the aligned encoders, we were able to adapt the text encoder to the demands of placenta analysis tasks and more effectively guide the image encoder. This approach addresses the limitations inherent in using a frozen pre-trained text encoder and allows for simultaneous training of both encoders while preserving the specific knowledge of placenta features learned during training.

Cross-modal distillation

In our prior research, we used a pre-trained bidirectional encoder representations from transformers (BERT) model[80,107] as the text encoder for encoding pathology reports, leveraging its language understanding and reasoning capabilities. BERT is a language model that is capable of understanding the relationships between words in a given text, which makes it particularly useful for tasks that involve language understanding and reasoning. For example, a trained BERT model can recognize that the presence of meconium staining may affect the diagnosis of other placenta features, such as inflammation responses and chorioamnionitis. This capability comes from BERT's extensive training on diverse text data, including medical documents that discuss the relationships between various terms. However, neither the pathology report nor the placenta image contains all the necessary information to model these relationships. Therefore, the text encoder only serves to encode the pathology report, and it is not designed to reason about these relationships, as suggested by Shen et al.[108] To address this limitation, we distill the knowledge from the pre-trained BERT by guiding the contrastive loss. Specifically, we split the text encoder f_u into the alignment text encoder f_{uc} and the reasoning text encoder f_{ub}, which allows us to obtain $u_{ci} = g_{uc}(f_{uc}(ti)))$ and $u_{bi} = g_{ub}(f_{ub}(t_i)))$. We applied the feature recomposition[45] technique on u_{bi} to reduce feature suppression. Then, we modified (Equation 1) into

$$\ell_i^{(v\to u)} = -\log\frac{\exp\left(\langle v_i, u_{ci}\rangle/\tau\right)}{\sum_{k=1}^{N}\exp\left(\langle v_i, u_{ck}\rangle/\tau\right)} - \lambda_1 \log\frac{\exp\left(\langle v_i, u_{bi}\rangle/\tau\right)}{\sum_{k=1}^{N}\exp\left(\langle v_i, u_{bk}\rangle/\tau\right)}, \quad \text{(Equation 3)}$$

where λ_1 is a hyperparemeter. The first objective aligns the image features with the alignment text feature to ensure that we have a text encoder that co-evolves with the image encoder as the training progresses; the second objective aligns the image features with the frozen BERT encoder to ensure that the reasoning capability of the text encoder is retained.

By distilling knowledge from the BERT encoder, we aim to improve the reasoning capabilities of the model and use the captured relationships between placental features that are not explicitly present in the pathology report or placenta image. This approach allows us to leverage the BERT's robust contextual modeling capabilities and improve the model's accuracy in placenta analysis tasks.

Cross-modal retrieval

The pathology report can vary in length depending on the number of placental features identified by the pathologist. The CLIP text encoder, however, is designed to accept a maximum of 77 text tokens, which means that we inevitably have to truncate the reports to fit the encoder. This truncation potentially results in information loss and may impair the model's ability to accurately encode the relationships between the words within the report. Additionally, CLIP generally performs better with shorter textual inputs, as it was trained on brief text segments. To address the issue, we previously randomly rearranged the placental features at each iteration to ensure that all features were covered in the training stage. Nonetheless, this approach of rearrangement and truncation may inadvertently generate false positive feature matches, where different text features are matched to the same image feature at each iteration or where features from non-relevant parts of the images are used in the matching process, which can cause the model to learn spurious relationships. False positive samples can have a negative impact on the model's performance and accuracy, as they can lead to robustness issues or poor generalization to new inputs.

To address the issue of false positives, we propose a cross-modal retrieval method. Traditional pooling layers use average pooling or its variants, which are based solely on the image modality, to obtain a feature vector from the feature map. Our method, however, necessitates aligning textual features with specific regions of the feature map based on the textual content. For example, if a pathology report mentions the presence of meconium staining, then we should not expect the model to match the textual feature related to meconium with the region of the feature map that corresponds to the umbilical cord. Our cross-modal retrieval module is designed to enhance the alignment between textual and visual features in a way that minimizes false positives and improves the model's performance. Formally, let V be the image feature map and u the corresponding textual feature. We obtain the textual query Q_u, the image key K_v, and the image value V_v as follows:

$$Q_u = \mathrm{LN}(u^\top)W_Q, K_v = \mathrm{LN}(V)W_K, V_v = \mathrm{LN}(V)W_V, \quad \text{(Equation 4)}$$

where *LN* is the layer-norm layer and Ws are the learnable weights. Then, we obtain the image feature based on the query text as $v'_u = \mathrm{LN}(\mathrm{Attn}(Q_u, K_v, V_v))$ and perform a final projection and residual connection following Gorti et al.[109] to obtain $v_u = \mathrm{LN}(\mathrm{FC}(v'_u) + v'_u)$, where *FC* is a fully connected layer. Then, we update Equation 3 to

$$\begin{aligned}\ell_i^{(v\to u)} = & -\log\frac{\exp\left(\langle v_i, u_{ci}\rangle/\tau\right)}{\sum_{k=1}^{N}\exp\left(\langle v_i, u_{ck}\rangle/\tau\right)} \\ & -\lambda_1\log\frac{\exp\left(\langle v_i, u_{bi}\rangle/\tau\right)}{\sum_{k=1}^{N}\exp\left(\langle v_i, u_{bk}\rangle/\tau\right)} \\ & -\lambda_2\log\frac{\exp\left(\langle v_{ui}, u_{ci}\rangle/\tau\right)}{\sum_{k=1}^{N}\exp\left(\langle v_{ui}, u_{ck}\rangle/\tau\right)},\end{aligned} \quad \text{(Equation 5)}$$

where the λs are hyperparameters.

By enhancing the alignment between textual and visual features, we can reduce the number of false positive samples and guide the model to learn more meaningful relationships between placental features and pathology reports, thereby improving its robustness.

Image pre-processing

We applied the AI-PLAX algorithm[42] to the primary dataset to mask out the background of each image, aligning with methodologies used in previous research.

We employed our AI-SAM algorithm,[75] trained using the dataset described in AI-PLAX,[42] to mask the background of each image in the external validation dataset. The preference for AI-SAM over AI-PLAX in segmentation tasks stems from AI-PLAX's limited robustness to domain shifts[43] and AI-SAM's capability for interactive modifications, advantageous in application settings.

All images were resized to 512×384 pixels to preserve all content. For pre-training augmentation, we applied random adjustments of brightness and contrast by up to 20%, saturation and hue shifts by up to 5%, and random rotation by up to 180°. For fine-tuning or validations, we used no augmentation.

Pathology report pre-processing

We used a simple pre-processing procedure for the pathology reports. The reports were split by anomalies and stored as a set. Irrelevant text, such as standard descriptions and information about the pathologist, was removed using keyword matching. When training our PlacentaCLIP model, we performed bootstrap sampling[45] from the set and concatenated the sampled items into complete sentences.

Model implementation

The image encoder was a ResNet50,[78] and the text encoder was a transformer model.[79] The BERT model[80] used for cross-modal distillation was trained in a self-supervised manner[107] on the MEDLINE/PubMed corpus.[110]

Pre-training stage

Our model and training code were written in Python 3.10.6 and PyTorch 1.11.0. Pre-training was conducted for 30 epochs with a batch size of 64. We utilized the PyTorch implementation of the AdamW optimizer[111] with default settings combined with a cosine learning rate scheduler. The initial learning rate was set to 1.0×10^{-5}, with a weight decay of 0.2 and 10% warm-up steps.

Fine-tuning stage

Evaluations were performed using scikit-learn 1.0.2 and pingouin 0.5.4. Fine-tuning was conducted by encoding the images using the pre-trained image encoder and training a logistic regression model for each task using the scikit-learn package. We trained the model using five nearly balanced random splits. An example is shown in Table 4.

Robustness evaluation dataset generation

Common image artifacts in placenta photos

Various artifacts can be introduced during placenta photo capture, potentially negatively affecting the accurate identification of placental features. In this study, we considered four common artifact types: blood stains, glare, JPEG compression, and shadow, as shown in Figure 8. Blood stains are often present

Table 4. An example random split of the fine-tuning dataset

	Meconium (n = 1,400)	FIR (n = 669)	MIR (n = 1,419)	Chorioamnionitis (n = 886)	Sepsis (n = 340)
Fine-tuning	700/1,400 (50.0%)	334/669 (49.9%)	709/1,419 (50.0%)	443/886 (50.0%)	170/340 (50.0%)
Positive	349/700 (49.9%)	167/334 (50.0%)	345/709 (48.7%)	224/443 (50.6%)	86/170 (50.6%)
Negative	351/700 (50.1%)	167/334 (50.0%)	364/709 (51.3%)	219/443 (49.4%)	84/170 (49.4%)
Evaluation	700/1,400 (50.0%)	335/669 (50.1%)	710/1,419 (50.0%)	443/886 (50.0%)	170/340 (50.0%)
Positive	330/700 (47.1%)	145/335 (43.3%)	378/710 (53.2%)	215/443 (48.5%)	80/170 (47.1%)
Negative	370/700 (52.9%)	190/335 (56.7%)	332/710 (46.8%)	228/443 (51.5%)	90/170 (52.9%)

The other four random splits are similar in number of each cases. FIR, fetal inflammatory response; MIR, maternal inflammatory response.

on the placenta and can obscure important features. These stains are usually dark red with irregular shapes, making them difficult to differentiate from actual features. Glare, caused by direct or reflected bright light sources, is common on reflective surfaces such as the placenta and can distort features, complicating identification. Many photo-taking devices, such as mobile phones, use JPEG compression to reduce file sizes, which can remove high-frequency or detailed information. While this generally does not affect standard object recognition tasks, it can be detrimental to identifying fine placental features. Shadows are another common challenge, appearing when the light source is behind the camera or when other objects cause uneven lighting across the placenta. Computationally, we simulated blood stains with randomly placed dark red spatters, glare with randomly placed white spatters, and shadows using a combination of randomly generated polygons and ellipses with blurred edges.

Common image blurring in placenta photos

Accurate diagnosis of placental features relies on observing low-level patterns present in the placenta photographs. However, when an image is blurry, these critical details are frequently distorted or lost, posing challenges to the diagnostic process. In practical settings, blurry placenta photographs are common, arising from a multitude of factors. To evaluate the robustness of the model, we introduced common types of blur, as shown in Figure 9, which include defocus, motion, and zoom blur. Defocus blur can result from an improperly focused lens when adjusting the camera position to accommodate the placenta's size. Motion blur may occur if the photographs are captured while the camera is still in motion. In addition, altering the camera's proximity to the placenta or using the lens' zoom function to adjust the placenta's apparent size in the image without pausing to refocus can cause zoom blur.

Common exposure artifacts in placenta photos

The brightness of an image refers to the perceived luminance of an image. It can be affected by both ambient lighting conditions and specific camera settings. For example, a lower exposure setting yields a darker image, while a higher exposure setting produces a brighter one. Similarly, contrast refers to the range of differentiation between the darkest and lightest parts within an image. This can be affected by the lighting conditions at the time of capture and the camera settings. For example, a photograph taken under bright sunlight generally shows a higher contrast than one taken for the same subject on an overcast day. Finally, saturation refers to the intensity and vividness of colors in an image. This can again be affected by lighting conditions and camera settings. For example, a subject photographed under brighter lighting conditions tends to show higher color vibrancy.

To evaluate the robustness of the placenta analysis model against alterations in brightness, contrast, and saturation, we systematically manipulated these attributes in the placenta photographs. This was achieved through adjusting the brightness, contrast, and saturation levels in the original images using image processing techniques, resulting in a set of images with varying levels of these factors as shown in Figure 10.

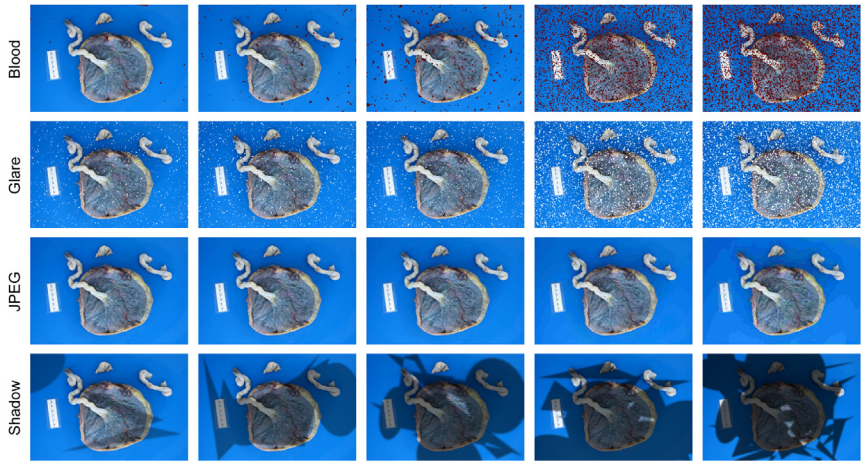


Figure 8. Examples of common image artifacts in placenta photos

The images from left to right are in the order of increasing corruption level.

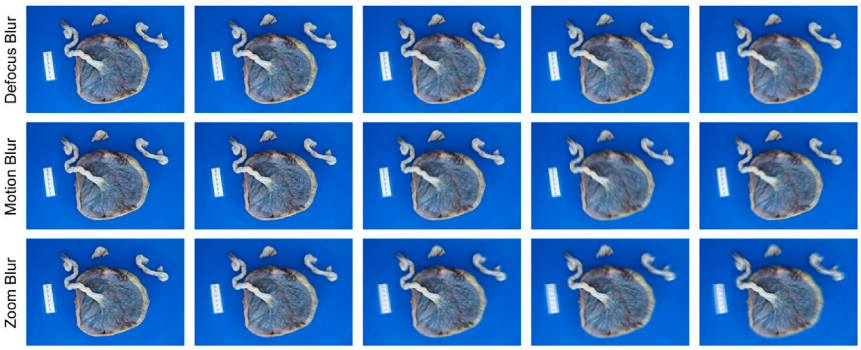


Figure 9. Examples of common image blur in placenta photos
The images from left to right are in the order of increasing corruption level.

Common white balance inaccuracies in placenta photos

White balance (WB) is the process of adjusting the color temperature of a photograph to eliminate color casts and accurately represent the colors in the image. The WB setting can drastically affect the appearance and diagnostic quality of a photo. When the color temperature of the light source does not match the camera's WB setting, the photo shows a color cast that distorts the original colors of the objects. For instance, an image taken under incandescent lighting might acquire a warm hue. WB is important in placenta analysis because the color of certain features in a placenta photo can be a key factor in accurate diagnosis. To evaluate the robustness of the model under different WB settings, we need to account for different color temperature preset options.

Various methods exist for adjusting WB, including using preset options, manual adjustment, or automatic correction. Modern digital cameras usually offer WB presets that cover common light sources, such as daylight, cloudy, tungsten light, and flash photography. Nonetheless, these presets are not always accurate and may require manual adjustment. Incorrect WB can also occur if it is not properly set or if lighting conditions change during a photo session (e.g., from sunny to overcast or from natural light to artificial light). To simulate the effects of incorrect WB on placenta analysis, we adopted the method proposed by Afifi and Brown,[83] which alters each placenta image to five different WB presets using two color profiles. This technique allowed us to evaluate the robustness of the model under different color casts commonly encountered in real-world scenarios. We followed the same approach and modified each placenta image to five different WB presets using two color profiles, as shown in Figure 11.

RESOURCE AVAILABILITY

Lead contact

Requests for information and resources used in this article should be addressed to the lead contact, James Wang (jwang@ist.psu.edu).

Materials availability

This study did not generate new unique reagents.

Data and code availability

Data collected in this study, including de-identified patient information, will be accessible upon reasonable request and subject to IRB approval. Access to the data requires the submission and approval of IRB protocols at both the originating and requesting institutions along with the execution of data

Figure 10. Examples of common exposure artifacts in placenta photos
The images from left to right are in the order of increasing corruption level.

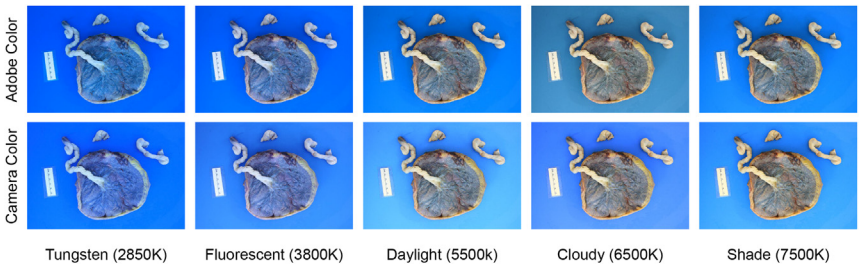


Figure 11. Examples of common WB inaccuracies in placenta photos

The images from left to right are in the order of increasing color temperature presets.

use agreements between the institutions. Currently, the data are owned by individual sites and shared with Penn State through established data use agreements.

Our source code is available on GitHub (https://github.com/ymp5078/PlacentaCLIP)[112] and has been archived on Zenodo.[113] Any additional information required to reanalyze the data reported in this paper is available from the lead contact upon request.

ACKNOWLEDGMENTS

Research reported in this publication was supported by the National Institute of Biomedical Imaging and Bioengineering of the National Institutes of Health (NIH) under award R01EB030130. Patient data used for training the algorithm were collected by Northwestern Memorial Hospital independent of the NIH's financial support. The external validation data came from a study supported by the National Institute of Child Health and Human Development of the NIH under awards R01HD112302 and K23AI138856 and the Burroughs Wellcome Fund/American Society of Tropical Medicine and Hygiene Postdoctoral Fellowship (ASTMH). The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH or ASTMH. This work used cluster computers at the National Center for Supercomputing Applications and the Pittsburgh Supercomputing Center through an allocation from the Advanced Cyberinfrastructure Coordination Ecosystem: Services & Support (ACCESS) program, which is supported by NSF grants 2138259, 2138286, 2138307, 2137603, and 2138296. The work also used the Extreme Science and Engineering Discovery Environment (XSEDE) under National Science Foundation grant ACI-1548562.

AUTHOR CONTRIBUTIONS

Y.P., J.A.G., A.D.G., and J.Z.W. contributed to the conceptualization and design of the experiments. Y.P. was responsible for methodology, software, formal analysis, visualization, and drafting of the manuscript. M.M., J.A.G., J.N., L.M.B., D.J.R., C.K.C., R.E.W., and A.D.G. contributed to data curation and the IRB protocol. A.D.G., J.Z.W., and J.A.G. acquired funding for the study. J.Z.W. secured high-performance computing resources. All authors contributed to the review and editing of the manuscript. Y.P. and J.Z.W. directly accessed and verified the raw data, and all authors had access to the data and had final responsibility for the decision to submit for publication.

DECLARATION OF INTERESTS

J.Z.W., A.D.G., and J.A.G. are named inventors on US patent 11,244,450, "Systems and Methods Utilizing Artificial Intelligence for Placental Assessment and Examination." It is assigned to The Penn State Research Foundation and Northwestern University. These interests do not influence the integrity of the research, and all efforts have been made to ensure that the research was conducted and presented in an unbiased manner.

DECLARATION OF GENERATIVE AI AND AI-ASSISTED TECHNOLOGIES IN THE WRITING PROCESS

During the preparation of this work, the authors used ChatGPT-4o in order to improve the readability of the manuscript. After using this tool/service, the authors reviewed and edited the content as needed and take full responsibility for the content of the published article.

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.patter.2024.101097.

Received: May 20, 2024
Revised: July 23, 2024
Accepted: October 23, 2024
Published: November 19, 2024

REFERENCES

1. Fitzgerald, E., Shen, M., Yong, H.E.J., Wang, Z., Pokhvisneva, I., Patel, S., O'Toole, N., Chan, S.-Y., Chong, Y.S., Chen, H., et al. (2023). Hofbauer cell function in the term placenta associates with adult cardiovascular and depressive outcomes. Nat. Commun. *14*, 7120. https://doi.org/10.1038/s41467-023-42300-8.
2. Ursini, G., Punzi, G., Chen, Q., Marenco, S., Robinson, J.F., Porcelli, A., Hamilton, E.G., Mitjans, M., Maddalena, G., Begemann, M., et al. (2018). Convergence of placenta biology and genetic risk for schizophrenia. Nat. Med. *24*, 792–801. https://doi.org/10.1038/s41591-018-0021-y.
3. Reis, A.S., Barboza, R., Murillo, O., Barateiro, A., Peixoto, E.P.M., Lima, F.A., Gomes, V.M., Dombrowski, J.G., Leal, V.N.C., Araujo, F., et al. (2020). Inflammasome activation and IL-1 signaling during placental malaria induce poor pregnancy outcomes. Sci. Adv. *6*, eaax6346. https://doi.org/10.1126/sciadv.aax6346.
4. Thornburg, K.L., and Marshall, N. (2015). The placenta is the center of the chronic disease universe. Am. J. Obstet. Gynecol. *213*, S14–S20. https://doi.org/10.1016/j.ajog.2015.08.030.
5. Barker, D.J.P., and Thornburg, K.L. (2013). Placental programming of chronic diseases, cancer and lifespan: a review. Placenta *34*, 841–845. https://doi.org/10.1016/j.placenta.2013.07.063.
6. Barker, D.J.P., Gelow, J., Thornburg, K., Osmond, C., Kajantie, E., and Eriksson, J.G. (2010a). The early origins of chronic heart failure: impaired placental growth and initiation of insulin resistance in childhood. Eur. J. Heart Fail. *12*, 819–825. https://doi.org/10.1093/eurjhf/hfq069.
7. Eriksson, J.G., Kajantie, E., Thornburg, K.L., Osmond, C., and Barker, D.J.P. (2011). Mother's body size and placental size predict coronary heart disease in men. Eur. Heart J. *32*, 2297–2303. https://doi.org/10.1093/eurheartj/ehr147.
8. Hutcheon, J.A., McNamara, H., Platt, R.W., Benjamin, A., and Kramer, M.S. (2012). Placental weight for gestational age and adverse perinatal outcomes. Obstet. Gynecol. *119*, 1251–1258. https://doi.org/10.1097/AOG.0b013e318253d3df.
9. Lema, G., Mremi, A., Amsi, P., Pyuza, J.J., Alloyce, J.P., Mchome, B., and Mlay, P. (2020). Placental pathology and maternal factors associated with stillbirth: An institutional based case-control study in northern tanzania. PLoS One *15*, e0243455. https://doi.org/10.1371/journal.pone.0243455.

10. Lou, S.K., Keating, S., Kolomietz, E., and Shannon, P. (2020). Diagnostic utility of pathological investigations in late gestation stillbirth: a cohort study. Pediatr. Dev. Pathol. *23*, 96–106. https://doi.org/10.1177/1093526619860353.

11. Levy, M., Alberti, D., Kovo, M., Schreiber, L., Volpert, E., Koren, L., Bar, J., and Weiner, E. (2020). Placental pathology in pregnancies complicated by fetal growth restriction: recurrence vs. new onset. Arch. Gynecol. Obstet. *301*, 1397–1404. https://doi.org/10.1007/s00404-020-05546-x.

12. Hauspurg, A., Redman, E.K., Assibey-Mensah, V., Tony Parks, W., Jeyabalan, A., Roberts, J.M., and Catov, J.M. (2018). Placental findings in non-hypertensive term pregnancies and association with future adverse pregnancy outcomes: a cohort study. Placenta *74*, 14–19. https://doi.org/10.1016/j.placenta.2018.12.008.

13. Holzman, C.B., Senagore, P., Xu, J., Dunietz, G.L., Strutz, K.L., Tian, Y., Bullen, B.L., Eagle, M., and Catov, J.M. (2021). Maternal risk of hypertension 7–15 years after pregnancy: clues from the placenta. BJOG An Int. J. Obstet. Gynaecol. *128*, 827–836. https://doi.org/10.1111/1471-0528.16498.

14. Curtin, W.M., Krauss, S., Metlay, L.A., and Katzman, P.J. (2007). Pathologic examination of the placenta and observed practice. Obstet. Gynecol. *109*, 35–41. https://doi.org/10.1097/01.AOG.0000437385.88715.4a.

15. Spencer, M.K., and Khong, T.Y. (2003). Conformity to guidelines for pathologic examination of the placenta: rates of submission and listing of clinical indications. Arch. Pathol. Lab Med. *127*, 205–207. https://doi.org/10.5858/2003-127-205-CTGFPE.

16. Taylor, L.A., Gallagher, K., Ott, K.A., and Gernand, A.D. (2021). How often is the placenta included in human pregnancy research? a rapid systematic review of the literature. Gates Open Res. *5*, 38. https://doi.org/10.12688/gatesopenres.13215.1.

17. Khong, T.Y., Mooney, E.E., Nikkels, P.G., Morgan, T.K., and Gordijn, S.J. (2019). Pathology of the Placenta: A Practical Guide (Springer). https://doi.org/10.1007/978-3-319-97214-5.

18. Brizuela, V., Cuesta, C., Bartolelli, G., Abdosh, A.A., Abou Malham, S., Assarag, B., Castro Banegas, R., Díaz, V., El-Kak, F., El Sheikh, M., et al. (2021). Availability of facility resources and services and infection-related maternal outcomes in the who global maternal sepsis study: a cross-sectional study. Lancet Global Health *9*, e1252–e1261. https://doi.org/10.1016/S2214-109X(21)00248-5.

19. Milton, R., Gillespie, D., Dyer, C., Taiyari, K., Carvalho, M.J., Thomson, K., Sands, K., Portal, E.A.R., Hood, K., Ferreira, A., et al. (2022). Neonatal sepsis and mortality in low-income and middle-income countries from a facility-based birth cohort: an international multisite prospective observational study. Lancet Global Health *10*, e661–e672. https://doi.org/10.1016/S2214-109X(22)00043-2.

20. Redline, R.W., Roberts, D.J., Parast, M.M., Ernst, L.M., Morgan, T.K., Greene, M.F., Gyamfi-Bannerman, C., Louis, J.M., Maltepe, E., Mestan, K.K., et al. (2023). Placental pathology is necessary to understand common pregnancy complications and achieve an improved taxonomy of obstetrical disease. Am. J. Obstet. Gynecol. *228*, 187–202. https://doi.org/10.1016/j.ajog.2022.08.010.

21. Vanea, C., Džigurski, J., Rukins, V., Dodi, O., Siigur, S., Salumäe, L., Meir, K., Parks, W.T., Hochner-Celnikier, D., Fraser, A., et al. (2024). Mapping cell-to-tissue graphs across human placenta histology whole slide images using deep learning with happy. Nat. Commun. *15*, 2710. https://doi.org/10.1038/s41467-024-46986-2.

22. Hutter, J., Harteveld, A.A., Jackson, L.H., Franklin, S., Bos, C., van Osch, M.J.P., O'Muircheartaigh, J., Ho, A., Chappell, L., Hajnal, J.V., et al. (2020). Perfusion and apparent oxygenation in the human placenta (PERFOX). Magn. Reson. Med. *83*, 549–560. https://doi.org/10.1002/mrm.27950.

23. Saini, B.S., Darby, J.R.T., Marini, D., Portnoy, S., Lock, M.C., Yin Soo, J., Holman, S.L., Perumal, S.R., Wald, R.M., Windrim, R., et al. (2021). An mri approach to assess placental function in healthy humans and sheep. J. Physiol. *599*, 2573–2602. https://doi.org/10.1113/JP281002.

24. Shchegolev, A.I., Tumanova, U.N., Lyapin, V.M., Kozlova, A.V., Bychenko, V.G., and Sukhikh, G.T. (2020). Complex method of CT and morphological examination of placental angioarchitechtonics. Bull. Exp. Biol. Med. *169*, 405–411. https://doi.org/10.1007/s10517-020-04897-4.

25. Aughwane, R., Schaaf, C., Hutchinson, J.C., Virasami, A., Zuluaga, M.A., Sebire, N., Arthurs, O.J., Vercauteren, T., Ourselin, S., Melbourne, A., and David, A.L. (2019). Micro-CT and histological investigation of the spatial pattern of feto-placental vascular density. Placenta *88*, 36–43. https://doi.org/10.1016/j.placenta.2019.09.014.

26. Beck, C., Gallagher, K., Taylor, L.A., Goldstein, J.A., Mithal, L.B., and Gernand, A.D. (2021). Chorioamnionitis and risk for maternal and neonatal sepsis: a systematic review and meta-analysis. Obstet. Gynecol. *137*, 1007–1022. https://doi.org/10.1097/AOG.0000000000004377.

27. Higgins, R.D., Saade, G., Polin, R.A., Grobman, W.A., Buhimschi, I.A., Watterberg, K., Silver, R.M., and Raju, T.N.K.; Chorioamnionitis Workshop Participants (2016). Evaluation and management of women and newborns with a maternal diagnosis of chorioamnionitis: summary of a workshop. Obstet. Gynecol. *127*, 426–436. https://doi.org/10.1097/AOG.0000000000001246.

28. Specktor-Fadida, B., Link-Sourani, D., Ferster-Kveller, S., Ben-Sira, L., Miller, E., Ben-Bashat, D., and Joskowicz, L. (2021). A bootstrap self-training method for sequence transfer: state-of-the-art placenta segmentation in fetal MRI Proc. Uncertainty for Safe Utilization of Machine Learning in Medical Imaging, and Perinatal Imaging. In Placental and Preterm Image Analysis (Springer), pp. 189–199. https://doi.org/10.1007/978-3-030-87735-4_18.

29. Pietsch, M., Ho, A., Bardanzellu, A., Zeidan, A.M.A., Chappell, L.C., Hajnal, J.V., Rutherford, M., and Hutter, J. (2021). APPLAUSE: Automatic Prediction of PLAcental health via U-net Segmentation and statistical Evaluation. Med. Image Anal. *72*, 102145. https://doi.org/10.1016/j.media.2021.102145.

30. Wang, Y., Li, Y.-Z., Lai, Q.-Q., Li, S.-T., and Huang, J. (2022). RU-Net: An improved U-Net placenta segmentation network based on ResNet. Comput. Methods Progr. Biomed. *227*, 1–7. https://doi.org/10.1016/j.cmpb.2022.107206.

31. Asadpour, V., Puttock, E.J., Getahun, D., Fassett, M.J., and Xie, F. (2023). Automated placental abruption identification using semantic segmentation, quantitative features, SVM, ensemble and multi-path CNN. Heliyon *9*, e13577. https://doi.org/10.1016/j.heliyon.2023.e13577.

32. Khodaee, A., Grynspan, D., Bainbridge, S., Ukwatta, E., and Chan, A.D. (2022). Automatic placental distal villous hypoplasia scoring using a deep convolutional neural network regression model. In Proc. IEEE Instrum. Meas. Technol. Conf. (IEEE), pp. 1–5. https://doi.org/10.1109/I2MTC48687.2022.9806589.

33. Dormer, J.D., Villordon, M., Shahedi, M., Leitch, K., Do, Q.N., Xi, Y., Lewis, M.A., Madhuranthakam, A.J., Herrera, C.L., Spong, C.Y., et al. (2022). CascadeNet for hysterectomy prediction in pregnant women due to placenta accreta spectrum. Proc. SPIE *12032*, 120320N. https://doi.org/10.1117/12.2611580.

34. Mobadersany, P., Cooper, L.A.D., and Goldstein, J.A. (2021). GestAltNet: Aggregation and attention to improve deep learning of gestational age from placental whole-slide images. Lab. Invest. *101*, 942–951. https://doi.org/10.1038/s41374-021-00579-5.

35. Sun, H., Jiao, J., Ren, Y., Guo, Y., and Wang, Y. (2023a). Multimodal fusion model for classifying placenta ultrasound imaging in pregnancies with hypertension disorders. Pregnancy Hypertens. *31*, 46–53. https://doi.org/10.1016/j.preghy.2022.12.003.

36. Ye, Z., Xuan, R., Ouyang, M., Wang, Y., Xu, J., and Jin, W. (2022). Prediction of placenta accreta spectrum by combining deep learning and radiomics using T2WI: A multicenter study. Abdom. Radiol. *47*, 4205–4218. https://doi.org/10.1007/s00261-022-03673-4.

37. Gupta, K., Balyan, K., Lamba, B., Puri, M., Sengupta, D., and Kumar, M. (2022). Ultrasound placental image texture analysis using artificial intelligence to predict hypertension in pregnancy. J. Matern. Fetal Neonatal Med. *35*, 5587–5594. https://doi.org/10.1080/14767058.2021.1887847.

38. Yampolsky, M., Salafia, C.M., Shlakhter, O., Haas, D., Eucker, B., and Thorp, J. (2008). Modeling the variability of shapes of a human placenta. Placenta *29*, 790–797. https://doi.org/10.1016/j.placenta.2008.06.005.

39. Haeussner, E., Schmitz, C., Von Koch, F., and Frank, H.-G. (2013). Birth weight correlates with size but not shape of the normal human placenta. Placenta *34*, 574–582. https://doi.org/10.1016/j.placenta.2013.04.011.

40. Ernst, L.M., Minturn, L., Huang, M.H., Curry, E., and Su, E.J. (2013). Gross patterns of umbilical cord coiling: correlations with placental histology and stillbirth. Placenta *34*, 583–588. https://doi.org/10.1016/j.placenta.2013.04.002.

41. Horikoshi, Y., Yaguchi, C., Furuta-Isomura, N., Itoh, T., Kawai, K., Oda, T., Matsumoto, M., Kohmura-Kobayashi, Y., Tamura, N., Uchida, T., et al. (2020). Gross appearance of the fetal membrane on the placental surface is associated with histological chorioamnionitis and neonatal respiratory disorders. PLoS One *15*, e0242579. https://doi.org/10.1371/journal.pone.0242579.

42. Chen, Y., Zhang, Z., Wu, C., Davaasuren, D., Goldstein, J.A., Gernand, A.D., and Wang, J.Z. (2020). AI-PLAX: AI-based placental assessment and examination using photos. Comput. Med. Imag. Graph. *84*, 101744. https://doi.org/10.1016/j.compmedimag.2020.101744.

43. Zhang, Z., Davaasuren, D., Wu, C., Goldstein, J.A., Gernand, A.D., and Wang, J.Z. (2020). Multi-region saliency-aware learning for cross-domain placenta image segmentation. Pattern Recogn. Lett. *140*, 165–171. https://doi.org/10.1016/j.patrec.2020.10.004.

44. Pan, Y., Gernand, A.D., Goldstein, J.A., Mithal, L., Mwinyelle, D., and Wang, J.Z. (2022). Vision-language contrastive learning approach to robust automatic placenta analysis using photographic images. In Proc. Int. Conf. Med. Image Comput. Comput. Assist. Interv. (Springer), pp. 707–716. https://doi.org/10.1007/978-3-031-16437-8_68.

45. Pan, Y., Cai, T., Mehta, M., Gernand, A.D., Goldstein, J.A., Mithal, L., Mwinyelle, D., Gallagher, K., and Wang, J.Z. (2023a). Enhancing automatic placenta analysis through distributional feature recomposition in vision-language contrastive learning. In Proc. Int. Conf. Med. Image Comput. Comput. Assist. Interv. (Springer), pp. 116–126. https://doi.org/10.1007/978-3-031-43987-2_12.

46. Zhang, Y., Jiang, H., Miura, Y., Manning, C.D., and Langlotz, C.P. (2022). Contrastive learning of medical visual representations from paired images and text. In Mach. Learn. for Healthc. Conf. (PMLR), pp. 2–25. https://openreview.net/forum?id=T4gXBOXoIUr.

47. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al. (2021). Learning transferable visual models from natural language supervision. In Proc. Int. Conf. Mach. Learn. (PMLR), pp. 8748–8763. https://proceedings.mlr.press/v139/radford21a.

48. Wen, K., Xia, J., Huang, Y., Li, L., Xu, J., and Shao, J. (2021). COOKIE: Contrastive cross-modal knowledge sharing pre-training for vision-language representation. In Proc. IEEE Comput. Soc. Int. Conf. Comput. Vis. (IEEE), pp. 2208–2217. https://doi.org/10.1109/ICCV48922.2021.00221.

49. Bakkali, S., Ming, Z., Coustaty, M., Rusiñol, M., and Terrades, O.R. (2023). VLCDoC: Vision-language contrastive pre-training model for cross-modal document classification. Pattern Recogn. *139*, 1–11. https://doi.org/10.1016/j.patcog.2023.109419.

50. Dong, X., Bao, J., Zheng, Y., Zhang, T., Chen, D., Yang, H., Zeng, M., Zhang, W., Yuan, L., Chen, D., et al. (2023). MaskCLIP: Masked self-distillation advances contrastive language-image pretraining. In Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (IEEE), pp. 10995–11005. https://doi.org/10.1109/CVPR52729.2023.01058.

51. Zhang, P., Li, X., Hu, X., Yang, J., Zhang, L., Wang, L., Choi, Y., and Gao, J. (2021). Vinvl: Revisiting visual representations in vision-language models. In Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (IEEE), pp. 5579–5588. https://doi.org/10.1109/CVPR46437.2021.00553.

52. Sun, Z., Fang, Y., Wu, T., Zhang, P., Zang, Y., Kong, S., Xiong, Y., Lin, D., and Wang, J. (2024). Alpha-CLIP: A clip model focusing on wherever you want. In Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (IEEE), pp. 13019–13029. https://openaccess.thecvf.com/content/CVPR2024/html/Sun_Alpha-CLIP_A_CLIP_Model_Focusing_on_Wherever_You_Want_CVPR_2024_paper.html.

53. Boecking, B., Usuyama, N., Bannur, S., Castro, D.C., Schwaighofer, A., Hyland, S., Wetscherek, M., Naumann, T., Nori, A., Alvarez-Valle, J., et al. (2022). Making the Most of Text Semantics to Improve Biomedical Vision–Language Processing. In Proc. Eur. Conf. Comput. Vis. (Springer), pp. 1–21. https://doi.org/10.1007/978-3-031-20059-5_1.

54. Li, T., Fan, L., Yuan, Y., He, H., Tian, Y., Feris, R., Indyk, P., and Katabi, D. (2023a). Addressing feature suppression in unsupervised visual representations. In Proc. IEEE Comput. Soc. Winter Conf. on Applications of Comput. Vis. (IEEE), pp. 1411–1420. https://doi.org/10.1109/WACV56688.2023.00146.

55. Zhai, X., Mustafa, B., Kolesnikov, A., and Beyer, L. (2023). Sigmoid loss for language image pre-training. In Proc. IEEE Comput. Soc. Int. Conf. Comput. Vis. (IEEE), pp. 11975–11986. https://doi.org/10.1109/ICCV51070.2023.01100.

56. Cui, Q., Zhou, B., Guo, Y., Yin, W., Wu, H., Yoshie, O., and Chen, Y. (2022). Contrastive vision-language pre-training with limited resources. In Proc. Eur. Conf. Comput. Vis. (Springer), pp. 236–253. https://doi.org/10.1007/978-3-031-20059-5_14.

57. Jia, C., Yang, Y., Xia, Y., Chen, Y.-T., Parekh, Z., Pham, H., Le, Q., Sung, Y.-H., Li, Z., and Duerig, T. (2021). Scaling up visual and vision-language representation learning with noisy text supervision. In Proc. Int. Conf. Mach. Learn. (PMLR), pp. 4904–4916. https://proceedings.mlr.press/v139/jia21b.

58. Sun, Q., Fang, Y., Wu, L., Wang, X., and Cao, Y. (2023b). EVA-CLIP: Improved training techniques for clip at scale. Preprint at arXiv. https://doi.org/10.48550/arXiv.2303.15389.

59. Wu, W., Sun, Z., Song, Y., Wang, J., and Ouyang, W. (2024). Transferring vision-language models for visual recognition: A classifier perspective. Int. J. Comput. Vis. *132*, 392–409. https://doi.org/10.1007/s11263-023-01876-w.

60. Vasu, P.K.A., Pouransari, H., Faghri, F., Vemulapalli, R., and Tuzel, O. (2024). MobileCLIP: Fast image-text models through multi-modal reinforced training. In Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (IEEE), pp. 15963–15974. https://openaccess.thecvf.com/content/CVPR2024/html/Vasu_MobileCLIP_Fast_Image-Text_Models_through_Multi-Modal_Reinforced_Training_CVPR_2024_paper.html.

61. Li, R., Kim, D., Bhanu, B., and Kuo, W. (2023b). RECLIP: Resource-efficient CLIP by training with small images. Trans. Mach. Learn. Res. https://openreview.net/forum?id=Ufc5cWhHko.

62. Chen, Y., Qi, X., Wang, J., and Zhang, L. (2023). DisCo-CLIP: A distributed contrastive loss for memory efficient clip training. In Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (IEEE), pp. 22648–22657. https://doi.org/10.1109/CVPR52729.2023.02169.

63. Zhao, Z., Liu, Y., Wu, H., Li, Y., Wang, S., Teng, L., Liu, D., Li, X., Cui, Z., Wang, Q., et al. (2023). CLIP in medical imaging: A comprehensive survey. Preprint at arXiv. https://doi.org/10.48550/arXiv.2312.07353.

64. Salari, S., Rasoulian, A., Rivaz, H., and Xiao, Y. (2023). Towards multimodal anatomical landmark detection for ultrasound-guided brain tumor resection with contrastive learning. In Proc. Int. Conf. Med. Image Comput. Comput. Assist. Interv. (Springer), pp. 668–678. https://doi.org/10.1007/978-3-031-43996-4_64.

65. Park, S., Lee, E.S., Shin, K.S., Lee, J.E., and Ye, J.C. (2024). Self-supervised multi-modal training from uncurated images and reports enables monitoring AI in radiology. Med. Image Anal. *91*, 103021. https://doi.org/10.1016/j.media.2023.103021.

66. Huang, S.-C., Shen, L., Lungren, M.P., and Yeung, S. (2021). GLoRIA: A multimodal global-local representation learning framework for

label-efficient medical image recognition. In Proc. IEEE Comput. Soc. Int. Conf. Comput. Vis. (IEEE), pp. 3922–3931. https://doi.org/10.1109/ICCV48922.2021.00391.

67. Müller, P., Kaissis, G., Zou, C., and Rueckert, D. (2022). Joint learning of localized representations from medical images and reports. In Eur. Conf. on Comput. Vis. (Springer), pp. 685–701. https://doi.org/10.1007/978-3-031-19809-0_39.
68. Cheng, P., Lin, L., Lyu, J., Huang, Y., Luo, W., and Tang, X. (2023). PRIOR: Prototype representation joint learning from medical images and reports. In Proc. IEEE Comput. Soc. Int. Conf. Comput. Vis. (IEEE), pp. 21361–21371. https://doi.org/10.1109/ICCV51070.2023.01953.
69. Zhang, K., Yang, Y., Yu, J., Jiang, H., Fan, J., Huang, Q., and Han, W. (2024). Multi-task paired masking with alignment modeling for medical vision-language pre-training. IEEE Trans. Multimed. *26*, 4706–4721. https://doi.org/10.1109/TMM.2023.3325965.
70. Zhang, X., Wu, C., Zhang, Y., Xie, W., and Wang, Y. (2023b). Knowledge-enhanced visual-language pre-training on chest radiology images. Nat. Commun. *14*, 4542. https://doi.org/10.1038/s41467-023-40260-7.
71. Wang, Z., Liu, C., Zhang, S., and Dou, Q. (2023). Foundation model for endoscopy video analysis via large-scale self-supervised pre-train. In Proc. Int. Conf. Med. Image Comput. Comput. Assist. Interv. (Springer), pp. 101–111. https://doi.org/10.1007/978-3-031-43996-4_10.
72. Wu, C., Zhang, X., Zhang, Y., Wang, Y., and Xie, W. (2023). Medklip: Medical knowledge enhanced language-image pre-training. In Proc. IEEE Comput. Soc. Int. Conf. Comput. Vis. (IEEE), pp. 21372–21383. https://doi.org/10.1109/ICCV51070.2023.01954.
73. Gernand, A.D., Goldstein, J., Wang, J.Z., Gallagher, K., and Walker, R.E. (2023). Placental Imaging Protocol (OSF). https://doi.org/10.17605/OSF.IO/9E6NM.
74. Khong, T.Y., Mooney, E.E., Ariel, I., Balmus, N.C.M., Boyd, T.K., Brundler, M.-A., Derricott, H., Evans, M.J., Faye-Petersen, O.M., Gillan, J.E., et al. (2016). Sampling and definitions of placental lesions: Amsterdam placental workshop group consensus statement. Arch. Pathol. Lab Med. *140*, 698–713. https://doi.org/10.5858/arpa.2015-0225-CC.
75. Pan, Y., Zhang, S., Gernand, A.D., Goldstein, J.A., and Wang, J.Z. (2023b). AI-SAM: Automatic and interactive segment anything model. Preprint at arXiv. https://doi.org/10.48550/arXiv.2312.03119.
76. Madrid, L., Alemu, A., Seale, A.C., Oundo, J., Tesfaye, T., Marami, D., Yigzaw, H., Ibrahim, A., Degefa, K., Dufera, T., et al. (2023). Causes of stillbirth and death among children younger than 5 years in eastern hararghe, ethiopia: a population-based post-mortem study. Lancet Global Health *11*, e1032–e1040. https://doi.org/10.1016/S2214-109X(23)00211-5.
77. Dhaded, S.M., Saleem, S., Goudar, S.S., Tikmani, S.S., Hwang, K., Guruprasad, G., Aradhya, G.H., Kusagur, V.B., Patil, L.G.C., Yogeshkumar, S., et al. (2022). The causes of preterm neonatal deaths in india and pakistan (PURPOSe): a prospective cohort study. Lancet Global Health *10*, e1575–e1581. https://doi.org/10.1016/S2214-109X(22)00384-9.
78. He, K., Zhang, X., Ren, S., and Sun, J. (2016). Deep residual learning for image recognition. In Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (IEEE), pp. 770–778. https://doi.org/10.1109/CVPR.2016.90.
79. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., and Polosukhin, I. (2017). Attention is all you need. Adv. Neural Inf. Process. Syst. *30*, 6000–6010. https://doi.org/10.5555/3295222.3295349.
80. Kenton, J.D.M.-W.C., and Toutanova, L.K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In Proc. NAACL-HLT (ACL), pp. 4171–4186. https://doi.org/10.18653/V1/N19-1423.
81. Camm, E., Wong, G., Pan, Y., Wang, J., Goldstein, J., Arcot, A., Murphy, C., Hansji, H., Mangwiro, Y., Saffery, R., et al. (2024). Assessment of an AI-based tool for population-wide collection of placental morphological data. Eur. J. Obstet. Gynecol. Reprod. Biol. *299*, 110–117. https://doi.org/10.1016/j.ejogrb.2024.05.043.
82. Hendrycks, D., and Dietterich, T. (2018). Benchmarking neural network robustness to common corruptions and perturbations. In Proc. Int. Conf. Learn. Represent https://openreview.net/forum?id=HJz6tiCqYm&hl=es.
83. Afifi, M., and Brown, M. (2019). What else can fool deep learning? addressing color constancy errors on deep neural network performance. In Proc. IEEE Comput. Soc. Int. Conf. Comput. Vis. (IEEE), pp. 243–252. https://doi.org/10.1109/ICCV.2019.00033.
84. Brier, G.W. (1950). Verification of forecasts expressed in terms of probability. Mon. Weather Rev. *78*, 1–3. https://doi.org/10.1175/1520-0493(1950)078<0001:VOFEIT>2.0.CO;2.
85. Benjamini, Y., and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. J. Roy. Stat. Soc.: Ser. Bibliogr. *57*, 289–300. https://doi.org/10.1111/j.2517-6161.1995.tb02031.x.
86. American College of Obstetricians and Gynecologists' Committee on Clinical Consensus-Obstetrics, Gantt, A., Society for Maternal-Fetal Medicine, Metz, T.D., Kuller, J.A., Louis, J.M., Society for Maternal-Fetal Medicine, Turrentine, M.A.; American College of Obstetricians and Gynecologists, and Cahill, A.G. (2023). Obstetric care consensus# 11, pregnancy at age 35 years or older. Am. J. Obstet. Gynecol. *228*, B25–B40. https://doi.org/10.1016/j.ajog.2022.07.022.
87. American College of Obstetricians and Gynecologists (2013). ACOG committee opinion No 579: Definition of term pregnancy. Obstet. Gynecol. *122*, 1139–1140. https://doi.org/10.1097/01.AOG.0000437385.88715.4a.
88. Health Level Seven International. Health level seven international. https://www.hl7.org/.
89. Alwasel, S.H., Harrath, A.-H., Aljarallah, J.S., Abotalib, Z., Osmond, C., Al Omar, S.Y., Thornburg, K., and Barker, D.J.P. (2013). The velocity of fetal growth is associated with the breadth of the placental surface, but not with the length. Am. J. Hum. Biol. *25*, 534–537. https://doi.org/10.1002/ajhb.22405.
90. Parks, W.T. (2015). Placental hypoxia: the lesions of maternal malperfusion. Semin. Perinatol. *39*, 9–19. https://doi.org/10.1053/j.semperi.2014.10.003.
91. Weiner, E., Mizrachi, Y., Grinstein, E., Feldstein, O., Rymer-Haskel, N., Juravel, E., Schreiber, L., Bar, J., and Kovo, M. (2016). The role of placental histopathological lesions in predicting recurrence of preeclampsia. Prenat. Diagn. *36*, 953–960. https://doi.org/10.1002/pd.4918.
92. Parks, W.T., and Catov, J.M. (2020). The placenta as a window to maternal vascular health. Obstet. Gynecol. Clin. *47*, 17–28. https://doi.org/10.1016/j.ogc.2019.10.001.
93. Chisholm, K.M., Heerema-McKenney, A., Tian, L., Rajani, A.K., Saria, S., Koller, D., and Penn, A.A. (2016). Correlation of preterm infant illness severity with placental histology. Placenta *39*, 61–69. https://doi.org/10.1016/j.placenta.2016.01.012.
94. Schlatterer, S.D., Murnick, J., Jacobs, M., White, L., Donofrio, M.T., and Limperopoulos, C. (2019). Placental pathology and neuroimaging correlates in neonates with congenital heart disease. Sci. Rep. *9*, 4137. https://doi.org/10.1038/s41598-019-40894-y.
95. Barker, D., Osmond, C., Grant, S., Thornburg, K.L., Cooper, C., Ring, S., and Davey-Smith, G. (2013a). Maternal cotyledons at birth predict blood pressure in childhood. Placenta *34*, 672–675. https://doi.org/10.1016/j.placenta.2013.04.019.
96. Barker, D.J.P., Osmond, C., Thornburg, K.L., Kajantie, E., and Eriksson, J.G. (2013b). The shape of the placental surface at birth and colorectal cancer in later life. Am. J. Hum. Biol. *25*, 566–568. https://doi.org/10.1002/ajhb.22409.
97. Barker, D.J.P., Osmond, C., Forsén, T.J., Thornburg, K.L., Kajantie, E., and Eriksson, J.G. (2013c). Foetal and childhood growth and asthma in adult life. Acta Paediatr. *102*, 732–738. https://doi.org/10.1016/j.placenta.2013.04.019.

98. Eriksson, J.G., Kajantie, E., Phillips, D.I.W., Osmond, C., Thornburg, K.L., and Barker, D.J.P. (2013). The developmental origins of chronic rheumatic heart disease. Am. J. Hum. Biol. *25*, 655–658. https://doi.org/10.1002/ajhb.22425.

99. Eriksson, J.G., Gelow, J., Thornburg, K.L., Osmond, C., Laakso, M., Uusitupa, M., Lindi, V., Kajantie, E., and Barker, D.J.P. (2012). Long-term effects of placental growth on overweight and body composition. Int. J. Paediatr *2012*, 324185. https://doi.org/10.1155/2012/324185.

100. Barker, D.J.P., Osmond, C., Thornburg, K.L., Kajantie, E., and Eriksson, J.G. (2013d). The intrauterine origins of hodgkin's lymphoma. Cancer Epidemiol. *37*, 321–323. https://doi.org/10.1016/j.canep.2013.01.004.

101. Barker, D.J.P., Larsen, G., Osmond, C., Thornburg, K.L., Kajantie, E., and Eriksson, J.G. (2012). The placental origins of sudden cardiac death. Int. J. Epidemiol. *41*, 1394–1399. https://doi.org/10.1093/ije/dys116.

102. Barker, D.J.P., Thornburg, K.L., Osmond, C., Kajantie, E., and Eriksson, J.G. (2010b). The surface area of the placenta and hypertension in the offspring in later life. Int. J. Dev. Biol. *54*, 525–530. https://doi.org/10.1387/ijdb.082760db.

103. PlacentaVision. PlacentaVision project. http://www.placentavision.com/.

104. U.S. Food and Drug Administration. How to determine if your product is a medical device. https://www.fda.gov/medical-devices/classify-your-medical-device/how-determine-if-your-product-medical-device.

105. Cubuk, E.D., Zoph, B., Mane, D., Vasudevan, V., and Le, Q.V. (2019). AutoAugment: Learning augmentation strategies from data. In Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (IEEE), pp. 113–123. https://doi.org/10.1109/CVPR.2019.00020.

106. De Francesco, D., Reiss, J.D., Roger, J., Tang, A.S., Chang, A.L., Becker, M., Phongpreecha, T., Espinosa, C., Morin, S., Berson, E., et al. (2023). Data-driven longitudinal characterization of neonatal health and morbidity. Sci. Transl. Med. *15*, eadc9854. https://doi.org/10.1126/scitranslmed.adc9854.

107. Google. BERT experts. https://www.kaggle.com/models/google/experts-bert/frameworks/tensorFlow2/variations/pubmed/versions/2.

108. Shen, S., Li, L.H., Tan, H., Bansal, M., Rohrbach, A., Chang, K.-W., Yao, Z., and Keutzer, K. (2021). How much can CLIP benefit vision-and-language tasks? In Proc. Int. Conf. Learn. Represent https://openreview.net/forum?id=zf_LI3HZWgy.

109. Gorti, S.K., Vouitsis, N., Ma, J., Golestan, K., Volkovs, M., Garg, A., and Yu, G. (2022). X-pool: Cross-modal language-video attention for text-video retrieval. In Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (IEEE), pp. 5006–5015. https://doi.org/10.1109/CVPR52688.2022.00495.

110. National Institutes of Health. Medline/PubMed corpus. https://www.nlm.nih.gov/databases/download/pubmed_medline.html.

111. Loshchilov, I., and Hutter, F. (2018). Decoupled weight decay regularization. In Proc. Int. Conf. Learn. Represent https://openreview.net/forum?id=Bkg6RiCqY7.

112. Pan, Y. (2024a). PlacentaCLIP Training and Evaluation (GitHub Repository). https://github.com/ymp5078/PlacentaCLIP.

113. Pan, Y. (2024b). PlacentaCLIP Training and Evaluation (Zenodo). https://doi.org/10.5281/zenodo.11412540.

Patterns, Volume *5*

Supplemental information

Cross-modal contrastive learning

for unified placenta analysis using photographs

Yimu Pan, Manas Mehta, Jeffery A. Goldstein, Joseph Ngonzi, Lisa M. Bebell, Drucilla J. Roberts, Chrystalle Katte Carreon, Kelly Gallagher, Rachel E. Walker, Alison D. Gernand, and James Z. Wang

Supplemental Methods

In this section, we provide additional results to support the analysis of PlacentaCLIP.

Including Stage 1 as Negative Samples in MIR and FIR

Table S1 presents the performance of MIR and FIR when stage 1 is included as a negative case. In the main text, stage 1 is excluded from the metric computation. We observe that including stage 1 as negative increases the standard deviation of the model's results, while the mean performance remains largely unchanged. This outcome is expected, as stage 1 is more challenging to identify.

Analysis of Variance for Combined Factors of Introduced Artificial Corruptions

We conducted an Analysis of Variance (ANOVA) analysis considering up to three combined effects, applied in a fixed order of corruption. These combined effect results encompass 175 comparisons for each task. Given the large size of the table containing these comparisons, it is provided separately in the Supplemental Information, allowing readers to filter. Only the significant variables with a P-value < 0.05 are shown here. Calculating all interactions for the 10 proposed corruptions in all possible orders would result in $10! = 3,628,800$ comparisons, which would be impractical. All analyses are conducted on PlacentaCLIP, in contrast to the more robust PlacentaCLIP+, to better reveal the differences in robustness.

Overall, the application of individual artifacts consistently impacts model performance, but the additional application of artifacts rarely affects performance (6.1% to 18.2% of the time). The model demonstrates the highest robustness in the sepsis classification task, with only 6.1% of the combined variables (Table S6) showing a statistically significant effect on performance. Conversely, the model exhibits the least robustness in the meconium classification task, where 18.2% of the combined variables (Table S2) show a statistically significant effect. Performance in the meconium classification task (Table S2) is most impacted by the additional application of color-based artifacts, such as saturation. In contrast, performance in other tasks (Table S3, Table S4, Table S5, and Table S6) is less affected by color-based artifacts. The performance in chorioamnionitis classification is less affected by the additional application of brightness artifacts compared to FIR (Table S4) and MIR (Table S5).

Due to the complexity of analyzing combined effects, more insights are drawn from direct performance comparisons discussed in the main text.

Effect of Class Distribution on Model Performance Across Different Demographic Groups

We conducted an additional analysis on the percentage of positive samples for each placenta feature and clinical outcome across different races in the internal validation set to identify any biases that might contribute to the model performing better on the 'Unknown' group than on the 'White' group. From Table S7, we observe that the 'Unknown' group has a significantly higher positive sample rate than the 'White' group for sepsis. Since our model performs best on the sepsis classification task, this disparity in class distribution is a contributing factor to the performance difference.

Table S1: **The performance of PlacentCLIP on MIR and FIR with and without discarding stage 1**

	mAP	STD	AUC	STD
FIR w/o stage 1	80.17	1.92	84.97	0.77
FIR w/ stage 1	78.47	4.81	82.72	2.71
MIR w/o stage 1	77.84	0.98	77.89	0.42
MIR w/ stage 1	79.11	2.77	79.97	3.35

The standard deviation increases when stage 1 is included as negative samples but the mean performance did not change too much. STD: standard deviation. mAP: mean average precision. AUC: area under the receiver operating characteristic curve.

Table S2: **Significant variables in the combined ANOVA analysis for the effects of synthetic artifacts on the performance of meconium**

Variables	SS	DoF	F	P-value
Blood	1064.381	1.000	351.344	0.000
Glare	179.239	1.000	59.165	0.000
Shadow	2238.017	1.000	738.753	0.000
Defocus blur	15374.586	1.000	5075.035	0.000
Motion blur	3786.043	1.000	1249.744	0.000
Zoom blur	704.864	1.000	232.670	0.000
Contrast	487.091	1.000	160.785	0.000
Brightness	2102.790	1.000	694.115	0.000
Saturation	1250.147	1.000	412.664	0.000
JPEG	7500.377	1.000	2475.818	0.000
Blood : Shadow	15.125	1.000	4.993	0.026
Blood : Defocus blur	13.441	1.000	4.437	0.036
Blood : Motion blur	14.585	1.000	4.815	0.029
Blood : Saturation	35.000	1.000	11.553	0.001
Blood : JPEG	62.768	1.000	20.719	0.000
Glare : Shadow	13.714	1.000	4.527	0.034
Glare : Defocus blur	45.330	1.000	14.963	0.000
Glare : Motion blur	56.777	1.000	18.742	0.000
Glare : Zoom blur	12.141	1.000	4.008	0.046
Glare : Brightness	20.076	1.000	6.627	0.010
Glare : JPEG	30.888	1.000	10.196	0.001
Shadow : Zoom blur	16.664	1.000	5.501	0.019
Shadow : Contrast	47.170	1.000	15.570	0.000
Shadow : Brightness	170.366	1.000	56.236	0.000
Shadow : JPEG	38.234	1.000	12.621	0.000
Defocus blur : Motion blur	386.506	1.000	127.583	0.000
Defocus blur : Zoom blur	50.099	1.000	16.537	0.000
Defocus blur : Contrast	21.960	1.000	7.249	0.007
Defocus blur : JPEG	186.820	1.000	61.668	0.000
Motion blur : JPEG	83.677	1.000	27.621	0.000
Zoom blur : Saturation	12.841	1.000	4.239	0.040
Contrast : Brightness	13.016	1.000	4.297	0.039
Blood : Defocus blur : Brightness	14.621	1.000	4.826	0.028
Blood : Defocus blur : JPEG	38.493	1.000	12.706	0.000
Blood : Motion blur : JPEG	12.135	1.000	4.006	0.046
Blood : Zoom blur : JPEG	11.805	1.000	3.897	0.049
Glare : Shadow : Saturation	13.342	1.000	4.404	0.036
Glare : Motion blur : JPEG	15.246	1.000	5.032	0.025
Glare : Saturation : JPEG	13.182	1.000	4.351	0.037
Shadow : Defocus blur : Brightness	43.748	1.000	14.441	0.000

Overall, the individual application of artifacts consistently produces a significant effect on model performance. However, the additional application of artifacts less frequently affects model performance (18.2% of the time). The performance of meconium is particularly affected by color artifacts such as brightness, contrast, and saturation. JPEG compression also causes an additional performance drop. DoF: degree of freedom. SS: sum of squares. F: F-statistics.

Table S3: **Significant variables in the combined ANOVA analysis for the effects of synthetic artifacts on the performance of chorioamnionitis**

Variables	SS	DoF	F	P-value
Blood	1373.483	1.000	236.825	0.000
Glare	601.677	1.000	103.745	0.000
Shadow	24.980	1.000	4.307	0.038
Defocus blur	6184.953	1.000	1066.450	0.000
Motion blur	2814.474	1.000	485.290	0.000
Zoom blur	710.379	1.000	122.488	0.000
Contrast	229.095	1.000	39.502	0.000
Brightness	1970.601	1.000	339.784	0.000
Saturation	109.970	1.000	18.962	0.000
JPEG	3944.526	1.000	680.141	0.000
Glare : Defocus blur	268.935	1.000	46.372	0.000
Glare : Motion blur	59.448	1.000	10.250	0.001
Glare : Zoom blur	24.644	1.000	4.249	0.040
Glare : JPEG	33.307	1.000	5.743	0.017
Shadow : Defocus blur	45.996	1.000	7.931	0.005
Shadow : JPEG	55.077	1.000	9.497	0.002
Defocus blur : Motion blur	138.841	1.000	23.940	0.000
Defocus blur : Zoom blur	27.204	1.000	4.691	0.031
Defocus blur : JPEG	209.837	1.000	36.182	0.000
Motion blur : Zoom blur	39.032	1.000	6.730	0.010
Motion blur : JPEG	22.394	1.000	3.861	0.050
Blood : Shadow : Motion blur	25.051	1.000	4.319	0.038
Blood : Defocus blur : JPEG	22.576	1.000	3.893	0.049
Glare : Defocus blur : Contrast	23.719	1.000	4.090	0.044

Overall, the individual application of artifacts always produces a significant effect on model performance. However, the additional application of artifacts less frequently affects model performance (8.5% of the time). The performance of chorioamnionitis tends to be affected by additional blur or JPEG compression. DoF: degree of freedom. SS: sum of squares. F: F-statistics.

Table S4: **Significant variables in the combined ANOVA analysis for the effects of synthetic artifacts on the performance of FIR**

Variables	SS	DoF	F	P-value
Blood	1614.241	1.000	333.964	0.000
Glare	535.509	1.000	110.789	0.000
Shadow	1150.898	1.000	238.105	0.000
Defocus blur	11910.097	1.000	2464.035	0.000
Motion blur	3292.658	1.000	681.206	0.000
Zoom blur	968.583	1.000	200.386	0.000
Contrast	350.373	1.000	72.487	0.000
Brightness	2976.847	1.000	615.869	0.000
Saturation	402.393	1.000	83.250	0.000
JPEG	7668.328	1.000	1586.471	0.000
Blood : Defocus blur	71.453	1.000	14.783	0.000
Blood : Brightness	48.186	1.000	9.969	0.002
Glare : Defocus blur	225.605	1.000	46.675	0.000
Glare : Motion blur	116.425	1.000	24.087	0.000
Glare : Zoom blur	32.331	1.000	6.689	0.010
Glare : JPEG	21.182	1.000	4.382	0.037
Shadow : Brightness	62.141	1.000	12.856	0.000
Shadow : Saturation	22.735	1.000	4.704	0.030
Defocus blur : Motion blur	56.735	1.000	11.738	0.001
Defocus blur : Brightness	19.432	1.000	4.020	0.045
Defocus blur : JPEG	20.588	1.000	4.259	0.039
Defocus blur : Motion blur : JPEG	30.999	1.000	6.413	0.012
Defocus blur : Brightness : JPEG	19.218	1.000	3.976	0.047

Overall, the individual application of artifacts always produces a significant effect on model performance. However, the additional application of artifacts less frequently affects model performance (7.9% of the time). The performance of FIR tends to be affected by additional blur or JPEG compression. DoF: degree of freedom. SS: sum of squares. F: F-statistics.

Table S5: **Significant variables in the combined ANOVA analysis for the effects of synthetic artifacts on the performance of MIR**

Variables	SS	DoF	F	P-value
Blood	2165.961	1.000	451.340	0.000
Glare	876.761	1.000	182.698	0.000
Shadow	392.768	1.000	81.844	0.000
Defocus blur	12153.796	1.000	2532.589	0.000
Motion blur	3499.102	1.000	729.138	0.000
Zoom blur	727.716	1.000	151.640	0.000
Contrast	119.325	1.000	24.865	0.000
Brightness	1610.976	1.000	335.693	0.000
Saturation	212.483	1.000	44.277	0.000
JPEG	8092.322	1.000	1686.266	0.000
Blood : Motion blur	19.020	1.000	3.963	0.047
Blood : JPEG	40.062	1.000	8.348	0.004
Glare : Shadow	37.054	1.000	7.721	0.006
Glare : Defocus blur	416.777	1.000	86.847	0.000
Glare : Motion blur	140.896	1.000	29.360	0.000
Glare : Zoom blur	28.145	1.000	5.865	0.016
Glare : Saturation	19.681	1.000	4.101	0.043
Shadow : Contrast	40.752	1.000	8.492	0.004
Shadow : Brightness	108.975	1.000	22.708	0.000
Shadow : JPEG	62.753	1.000	13.076	0.000
Defocus blur : Motion blur	201.530	1.000	41.995	0.000
Defocus blur : Zoom blur	26.709	1.000	5.566	0.019
Defocus blur : Brightness	25.252	1.000	5.262	0.022
Defocus blur : JPEG	628.988	1.000	131.067	0.000
Motion blur : JPEG	91.318	1.000	19.029	0.000
Defocus blur : Motion blur : Zoom blur	28.910	1.000	6.024	0.014

Overall, the individual application of artifacts always produces a significant effect on model performance, but additional applications of artifacts affect model performance less frequently (9.7% of the time). The performance of MIR tends to be affected by additional blur or JPEG compression. DoF: degree of freedom. SS: sum of squares. F: F-statistics.

Table S6: **Significant variables in the combined ANOVA analysis for the effects of synthetic artifacts on the performance of sepsis**

Variables	SS	DoF	F	P-value
Blood	617.963	1.000	101.651	0.000
Glare	886.889	1.000	145.887	0.000
Shadow	1431.548	1.000	235.480	0.000
Defocus blur	3247.595	1.000	534.207	0.000
Motion blur	813.714	1.000	133.850	0.000
Zoom blur	217.950	1.000	35.851	0.000
Contrast	388.370	1.000	63.884	0.000
Brightness	1512.863	1.000	248.855	0.000
Saturation	170.926	1.000	28.116	0.000
JPEG	576.612	1.000	94.849	0.000
Blood : Defocus blur	27.230	1.000	4.479	0.035
Glare : Defocus blur	145.333	1.000	23.906	0.000
Glare : Motion blur	41.600	1.000	6.843	0.009
Glare : JPEG	51.284	1.000	8.436	0.004
Shadow : JPEG	59.986	1.000	9.867	0.002
Defocus blur : Motion blur	79.073	1.000	13.007	0.000
Defocus blur : Brightness	50.342	1.000	8.281	0.004
Saturation : JPEG	27.838	1.000	4.579	0.033
Shadow : Defocus blur : Brightness	28.600	1.000	4.705	0.030
Defocus blur : Motion blur : Brightness	28.522	1.000	4.692	0.031

Overall, the individual application of artifacts always produces a significant effect on model performance, but additional applications of artifacts affect model performance less frequently (6.1% of the time). The performance of sepsis tends to be affected by additional blur or JPEG compression. DoF: degree of freedom. SS: sum of squares. F: F-statistics.

Table S7: **The percentage of positive samples for each placenta feature and clinical outcome across different races in the internal validation set**

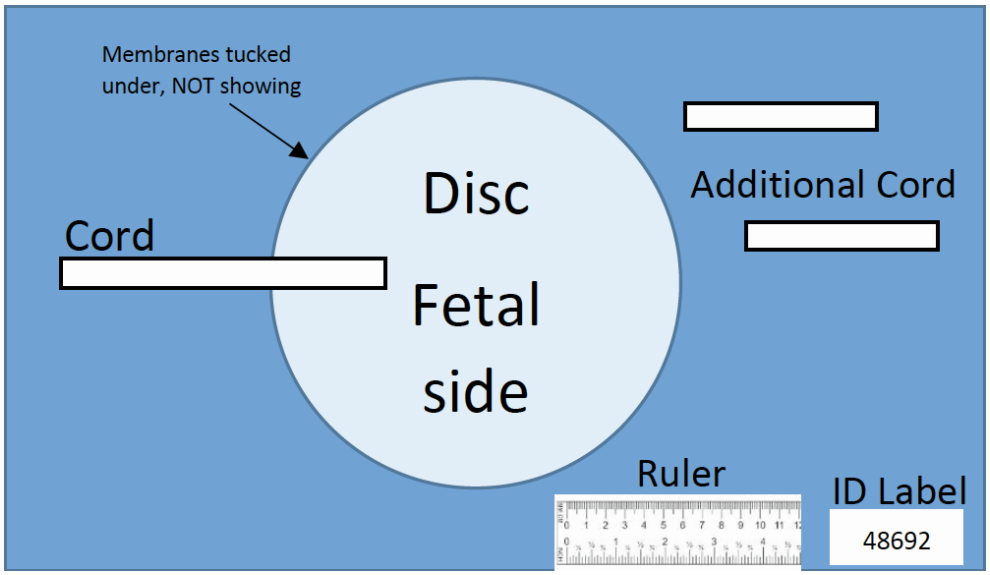
	Meconium	Chorioamnionitis	FIR	MIR	Sepsis
Asian	50.00%	57.14%	39.39%	61.04%	20.00%
Black or African American	45.35%	52.31%	50.00%	53.47%	18.18%
Other	49.46%	65.57%	52.00%	58.59%	23.53%
Unknown	40.43%	57.89%	39.13%	59.70%	25.00%
White	46.34%	39.66%	41.30%	49.45%	7.02%

FIR: fetal inflammatory response; MIR: maternal inflammatory response.

Data S1. Photo-taking procedure at Mbarara Regional Referral Hospital

Photographing the placenta – **FETAL** surface

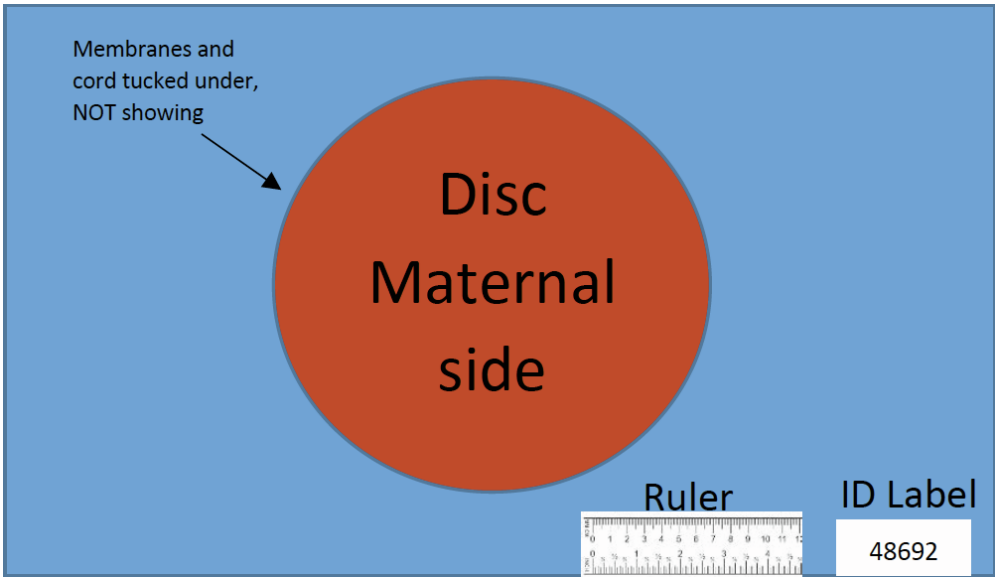
- Tuck membranes under disc so they are not showing
- **Lay cord <u>off</u> of placenta from the shortest edge** – start by laying it straight off the closest edge of the placenta
- Arrange the cord straight/curved around the disc (NOT curled up); **don't let cord touch the disc or touch itself**
- Wipe the placental surface to reduce glare (make sure there is no light reflecting on the placenta's shiny surface)
- Wipe blood and clots off disc and cord
- Wipe blood off blue board
- Include entire ruler in photo (bottom)
- Include ID in photo (bottom right)



Take fetal side photo – check that image is **in focus** with **no glare** and **no shadows** – take a second photo to be sure
Take additional, close-up photograph(s) of any visible lesions

Photographing the placenta – **MATERNAL** surface

- Turn the disc over
- Center disc and tuck membranes and cord under disc (NO cord segments should be visible)
- Wipe blood and clots off maternal disc surface
- Wipe blood off of blue background.
- Include all fragments of the placenta if disc is not whole
- Ensure ruler and ID number are still in place in the photo and are not touching disc or cord



Take maternal side photo – check that image is **in focus** with **no glare** and **no shadows** – take a second photo to be sure
Take additional, close-up photograph(s) of any visible lesions

CHECK your photos

Look at the back of the camera, find the button with the small arrow:

Press the left side of the large round button to look at the previous photo

Press the right side of the large round button to look at the next photo

CHECK to be sure:

1. The photo is clear, in focus, and not blurry
2. All portions of the placenta are entirely contained in the photo
3. These are **NOT** in the photo:
    a. Non-adherent blood clots
    b. Blood on the disc of blue background
    c. Hands
    d. Feet
    e. Clothing
    f. Tools
    g. Containers
    h. Wipes or paper towels
    i. Blood stains
    j. Motion
    k. Glare (reflection – use gauze to dry wet areas)
    l. Uneven lighting, shadows
3. The placenta and cord are not be touching
4. The cord is not touching itself
5. The entire ruler is visible in the photo
6. There is space between the items – blue background is visible in-between each item

TAKE MORE PHOTOGRAPHS if you see any of these problems, or if you're not sure (it is easier to delete them later)